University of Alberta

Regularization in Reinforcement Learning

by

Amir-massoud Farahmand

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Computing Science

Fall 2011
Edmonton, Alberta

Abstract

This thesis studies the reinforcement learning and planning problems that are modeled by a discounted Markov Decision Process (MDP) with a large state space and finite action space. We follow the value-based approach in which a function approximator is used to estimate the optimal value function. The choice of function approximator, however, is nontrivial, as it depends on both the number of data samples and the MDP itself. The goal of this work is to introduce flexible and statistically-efficient algorithms that find close to optimal policies for these problems without much prior information about them.

The recurring theme of this thesis is the application of the regularization technique to design value function estimators that choose their estimates from rich function spaces. We introduce regularization-based Approximate Value/Policy Iteration algorithms, analyze their statistical properties, and provide upper bounds on the performance loss of the resulted policy compared to the optimal one. The error bounds show the dependence of the performance loss on the number of samples, the capacity of the function space to which the estimated value function belongs, and some intrinsic properties of the MDP itself. Remarkably, the dependence on the number of samples in the task of policy evaluation is minimax optimal.

We also address the problem of automatic parameter-tuning of reinforcement learning/planning algorithms and introduce a complexity regularization-based model selection algorithm. We prove that the algorithm enjoys an oracle-like property and it may be used to achieve adaptivity: the performance is almost as good as the performance of the unknown best parameters.

Our two other contributions are used to analyze the aforementioned algorithms. First, we analyze the rate of convergence of the estimation error in regularized least-squares regression when the data is exponentially β-mixing. We prove that up to a logarithmic factor, the convergence rate is the same as the optimal minimax rate available for the i.i.d. case. Second, we attend to the question of how the errors at each iteration of the approximate policy/value iteration influence the quality of the resulting policy. We provide results that highlight some new aspects of these algorithms.

Acknowledgements

The process of getting a PhD starts much earlier than the time of entering graduate school. In my case, I can trace the influence of people and events for a quarter of century. I will not try to relay the path that led me to become interested in machine intelligence in general and reinforcement learning in particular, as it is a long story serving no benefit to the reader of this thesis. I would like, however, to express my gratitude to a few people who have directly or indirectly influenced my studies and myself in the past few years.

I would like to thank Csaba Szepesvári, my supervisor, for all his support and encouragement in the past five years. I learned from him that the rigorous mathematical treatment of machine learning problems can be both insightful and fun. He was the one who taught me how to write clearly. He has always been there to help me out when I got stuck in my research. Without him, this research could not be done. Thank you Csaba!

Martin Jägersand has always been supportive and kind to me. He helped me since well before my admission until the end of my education. Even though the research we have done together did not end up in this thesis, I learned a great deal about visual servoing and many other things from him.

I am thankful to Richard Sutton, Dale Schuurmans, Michael Bowling, Alexander Melnikov, Ivan Mizera, and Peter Bartlett for accepting to be on my committee and providing constructive feedback on my research. Rich's ideas on how AI should be approached have been inspiring and thought-provoking. This thesis may not explicitly reflect those ideas, but they are inscribed in my brain. Dale has been a great example of how one can teach clearly and ask smart questions. He is my role model of how a teacher should be.

During my PhD studies, I collaborated with several great researchers and benefitted from their wisdom. Other than my supervisors, Csaba and Martin, I have enjoyed working with Jean-Yves Audibert, Azad Shademan, Mohammad Ghavamzadeh, Shie Mannor, and Rémi Munos. Mohammad was not only my collaborator, but also an academic mentor and a good friend. And I happily recall our memorable research and non-research discussions during our coffee-breaks with Azad.

I am thankful to Mohammad Sadegh Abrishamian, my BS supervisor, for being such an aspiring teacher who made sure that I always aimed for my full potential. I would not be interested in reinforcement learning if I did not have Majid Nili Ahmadabadi as my MS supervisor. Our discussions are still among the best I have ever had. I am thankful to Babak Araabi for his encouraging optimism and kindness. And I wish we still had our dear Caro Lucas for he was a spring of ideas, inspiration, and humanity.

I learned a lot from the members of the RLAI, Vision, and Robotics labs. I greatly enjoyed our meetings, tea-time talks, and our retreats. Thank you for that. I also express my gratitude to the staff of our department, especially Edith Drummond and Lori Troop, for among other things taking care of all documents that I ever needed and all bills that had to be reimbursed. I am grateful to Doina Precup for offering me a Postdoc position and then waiting for a long time before I actually arrived!

This thesis is admittedly long. As a result, its proofreading was a challenge that not many people dared to face. Those brave individuals who read it are Csaba Szepesvári, Mohammad Ghavamzadeh, Shie Mannor, István Szita, Azad Shademan, and Frances Reilly. I appreciate their time and effort that greatly helped this thesis to be more readable. I am

specially thankful to Csaba who made sure that the thesis is technically sound and Frances who guided me on how to use the English language properly. Needless to say, all remaining typos and errors are my fault.

I enjoyed the friendship of many great people in the past few years. There is no enough space to mention all of them, so I only name a few whose friendship has been very important and dear to me, from almost the beginning of my stay in Edmonton to the moment: Alborz Geramifard, Amin Jorati, AmirAli Sharifi, Amir Pouyan Shiva, Arash Afkanpour, Azad Shademan, Babak Damavandi, Barnabas Poczos, David Lovi, Frances Reilly, Hamid Maei, Hootan Nakhost, Katayoon Navabi, Kiana Hajebi, Metanat HooshSadat, Mina Mahdavi, Maysam Heydari, Mohammad Ghavamzadeh, Mohammad Gheshlaghi Azar, Mohammad Shafie, Nader Damavandi, Neda Mirian, Niousha Bolandzadeh, Paymon Hamed Hosseini, Ramin Mehran, Reihaneh Rabbany, Reza Saddodin, Roshanak Nilchiani, Roozbeh Fazl, Saman Vaisipour, Siamak Ravanbakhsh, Sima Sajjadian, The Simpsons, Varun Grover, Yasin Abbasi-Yadkori, Yavar Naddaf, and several others. I am glad that I have met you.

My special thanks go to my dearest Frances whose love makes the world a much nicer place to be and my life a much happier experience to live.

Finally, I am indebted to my family for their unconditional love. I would not be where I am today without their encouragement and support. Thank you Delshad, Navid, Kian, Maman-joon, and Baabi.

Contents

List of Figures

List of Algorithms

List of Symbols

MDP

$\mathcal{X}$**:** State space.

$\mathcal{A}$**:** Action space – A finite set with cardinality $|\mathcal{A}|$.

$\mathcal{X} \times \mathcal{A}$**:** State-Action space.

$P : \mathcal{X} \times \mathcal{A} \to \mathcal{M}(\mathbb{R} \times \mathcal{X})$**:** Reward-transition probability kernel.

$P : \mathcal{X} \times \mathcal{A} \to \mathcal{M}(\mathcal{X})$**:** Transition probability kernel.

$\mathcal{R}$**:** Reward distribution.

γ**:** Discount factor.

π**:** Policy.

V^π**:** Value function for policy π.

Q^π**:** Action-value function for policy π.

V^***:** Optimal value function.

Q^***:** Optimal action-value function.

$\hat{\pi}(\cdot; Q)$**:** Greedy policy w.r.t. the action-value function Q.

T^π**:** Bellman operator.

T^***:** Bellman optimality operator.

$\hat{T}^\pi$**:** Empirical Bellman operator.

$\hat{T}^*$**:** Empirical Bellman optimality operator.

Probability and Samples

$\sigma_{\mathcal{X}}$**:** The σ-algebra defined on $\mathcal{X}$.

$\mathcal{M}(\mathcal{X})$**:** The set of all probability measures defined on $\sigma_{\mathcal{X}}$.

$X \sim P \in \mathcal{M}(\mathcal{X})$**:** X is a sample drawn from distribution P.

ρ**:** The performance evaluation distribution.

ν**:** The data sampling distribution underlying $\{(X_t, A_t)\}$.

$\nu_{\mathcal{X}}$**:** The data sampling distribution underlying $\{X_t\}$.

$\mathcal{D}_n = \{(X_1, A_1, R_1, X'_1), \ldots, (X_n, A_n, R_n, X'_n)\}$**:** Data samples used for RFQI, REG-LSTD, REG-BRM, and BErMin.

$\nu_1 \ll \nu_2$**:** ν_1 is dominated by ν_2, i.e., $\nu_2(A) = 0 \Rightarrow \nu_1(A) = 0$.

Function Spaces

$B(\mathcal{X})$**:** The space of bounded measurable function w.r.t. the σ-algebra $\sigma_{\mathcal{X}}$.

$B(\mathcal{X}, L)$**:** The space of bounded measurable functions w.r.t. the σ-algebra $\sigma_{\mathcal{X}}$ with the bound $L < \infty$.

$\mathcal{F}$**:** A subset of measurable functions $\mathcal{X} \to \mathbb{R}$.

$\mathcal{F}^{|\mathcal{A}|}$**:** A subset of vector-valued measurable functions $\mathcal{X} \times \mathcal{A} \to \mathbb{R}^{|\mathcal{A}|}$.

$\mathcal{H}$**:** reproducing kernel Hilbert space.

$\mathbb{W}^k(\mathbb{R}^d)$**:** Sobolev space $\mathbb{W}^{k,2}(\mathbb{R}^d)$.

$\|f\|_{p,\nu}$**:** $\left(\int_{\mathcal{X}\times\mathcal{A}} |f(x,a)|^p d\nu(x,a)\right)^{1/p}$.

$\|f\|_{p,\nu,n}$**:** $\left(\frac{1}{n}\sum_{t=1}^n f^p(X_t, A_t)\right)^{1/p}$, $(X_t, A_t) \sim \nu$. (Simplified: $\|\cdot\|_{p,n}$)

$L_p(\mathcal{X} \times \mathcal{A}, \nu)$**:** The set of all measurable functions on $\mathcal{X} \times \mathcal{A}$ with finite $\|f\|_{p,\nu}$.

$J(f)$**:** Regularizer/Penalizer of f.

$\Pi_\nu = \Pi_{\nu,\mathcal{F}^{|\mathcal{A}|}}$**:** Projection operator onto $\mathcal{F}^{|\mathcal{A}|}$, i.e., $\Pi_{\nu,\mathcal{F}^{|\mathcal{A}|}} f \triangleq \operatorname{argmin}_{f' \in \mathcal{F}^{|\mathcal{A}|}} \|f' - f\|_\nu^2$ for $f \in B(\mathcal{X} \times \mathcal{A})$.

Capacity of Function Spaces

$\mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|_{p,\nu})$**:** The ε-covering number of $\mathcal{F}$ w.r.t. $\|\cdot\|_{p,\nu}$.

$\mathcal{N}_p(\varepsilon, \mathcal{F}, x_{1:n})$**:** The empirical ε-covering number of $\mathcal{F}$ w.r.t. the norm $\|\cdot\|_{p,n}$.

$\mathcal{M}(\varepsilon, \mathcal{F}, \|\cdot\|_{p,\nu})$**:** The ε-packing number of $\mathcal{F}$ w.r.t. $\|\cdot\|_{p,\nu}$.

Others

$\mathbb{N}$**:** The positive natural numbers $\{1, 2, 3, \dots\}$.

$\mathbb{N}_0$**:** The natural numbers $\{0, 1, 2, \dots\}$.

$\mathbb{R}$**:** Real numbers.

$\mathbb{I}_{\{\cdot\}}$**:** Indicator function – for an event E, $\mathbb{I}_{\{E\}} = 1$ if and only if E holds and $\mathbb{I}_{\{E\}} = 0$, otherwise.

$a \wedge b$**:** $\min\{a, b\}$

$a \vee b$**:** $\max\{a, b\}$

$\Theta(\cdot)$**:** $\Theta(f(n)) = \{g(n) : \exists c_1, c_2 > 0$ and n_0 s.t. $0 \le c_1 g(n) \le f(n) \le c_2 g(n)$ for $n \ge n_0\}$.

Δ_π**:** $\Delta_\pi : \mathcal{X} \times \mathcal{A} \to [0, 1]$ is a function with the property that $\sum_{a\in\mathcal{A}} \Delta_\pi(x, a) \le \varepsilon \le 1$ for all $x \in \mathcal{X}$.

Chapter 1

Introduction

Many real-world decision-making problems can be described as either a **Reinforcement Learning (RL)** or a **Planning** problem. More often than not, these sequential decision-making problems have large state spaces, and to efficiently solve them one usually has to rely on the use of some **function approximation** method.[1] The appropriate choice of function approximation for a given problem, however, is far from trivial. The suitable choice depends on many factors including the problem itself and the way one interacts with it. Different problems call for different function approximators in a way that is not easy – if not impossible at all – to know prior to solving the problem itself. The high-level goal of this thesis is to introduce and analyze flexible and statistically-efficient methods that can solve RL/Planning problems with large state spaces.

We use the following household humanoid robot example as an instance of a sequential decision-making problem with large state space. Nevertheless, we do not focus on any specific application domain later, and our emphasis will be on theoretical studies.

The Household Humanoid Example

Imagine a humanoid robot [Kemp et al., 2008] that is responsible for running a household and interacting with humans. The robot can sense the external world through its stereo-vision cameras, microphones, and tactile sensors all around its body. Moreover, in order to handle delicate tasks such as grasping dishes and using stairs, it has a motor-rich body with tens of degrees of freedom. The goal of the designer is to develop an *"artificial mind"* (i.e., a decision-maker) that receives sensory inputs, and provides appropriate motor commands so that the robot can successfully complete the required tasks.

This problem is an instance of sequential decision-making problems. It is sequential because many tasks, such as preparing a meal or doing the laundry, have a temporal aspect and achieving them requires a well-planned sequence of actions. The robot also requires to deal with large state spaces. Consider the robot's sensory inputs such as its cameras that provide high-dimensional real-valued inputs. The decision-maker may summarize all these sensory inputs in an internal representation, which we informally call *state*, and then base its decision on the robot's current state. The size of the state space, however, might be huge if the state is supposed to represent the external world accurately – especially if the external world is unstructured and its description cannot be considerably compressed.

This example is merely an instance of sequential decision-making problems with large state spaces. Other fields of robotics, for example visual-servoing of manipulator arms [Chaumette and Hutchinson, 2008; Farahmand et al., 2007a, 2009c; Shademan et al., 2010] and mobile robots [Siciliano and Khatib, 2008, Part E] also require to solve similar problems. More generally, almost all control engineering problems are instances of sequential decision-making problems.

[1] We define *state space* and other related concepts in Chapter 2.

A flexible computational framework for solving sequential decision-making problems with large state spaces accompanied by a good theoretical understanding has far reaching applications. In addition to robotics and control engineering, researchers have found RL/Planning useful in finance and have applied it to problems such as optimized trade execution [Nevmyvaka et al., 2006] and to learning exercise policy for American options [Li et al., 2009]. Healthcare applications of reinforcement learning methods, and especially the dynamic treatment regime problem, are emerging [Pineau et al., 2007]. And finally, reinforcement learning algorithms have also been used to design automated players for games such as backgammon [Tesauro, 1994], Go [Silver et al., 2007], and Atari 2600 console games [Naddaf, 2010]. See Szita [2011] for a recent survey on the applications of RL in computer games.

1.1 Regularities and Adaptive Algorithms

How well can we solve any RL/Planning problem?

Negative results from the supervised learning theory suggest that "efficient" learning is hopeless for general classes of problems, see e.g., Theorem A.1 in Section A.1. The situation cannot be better for RL/Planning problems as they are supersets of regression problems. It is indeed impossible to design a universal RL/Planning method that works "efficiently" for *all* problems.

Fortunately, not all decision-making problems are equally difficult. If one finds some kind of structure or **regularity** in a given problem, he can solve the problem with much less effort. Examples of such regularities for sequential decision-making problems are the *smoothness* of the value function[2], the *sparsity* of the value function in a certain basis, or the input data lying close to *a low-dimensional manifold* (see Section A.2).[3]

Results from the supervised learning theory ensure that an algorithm benefitting from the regularities of the problem may perform reasonably well. Two key points deserve more emphasis. The first is that the *problem* itself must have some kind of regularity. For example, the value function should be smooth, or could be described by a few dimensions of the state space. Regularity is an *intrinsic* property of the problem. The second key point is that the algorithm should be capable to exploit the present regularities of the problem. For instance, a conventional K-Nearest Neighborhood-based algorithm cannot benefit from the smoothness of value function, and as a result its performance would be almost identical to when the algorithm is faced with a problem without such smoothness regularity.

A highly desirable requirement for any learning agent is to *adapt* to the intrinsic difficulty of the learning problem. Whenever the problem has certain regularities, we would like the agent to deliver a better solution with the same amount of data/interactions with the environment. If an agent can automatically exploit these regularities and provide a solution as if it knew the right regularity of the problem a priori, we call it an **adaptive** agent/algorithm.

To clarify the notions of regularity and adaptivity, consider a simple numerical analysis example: the problem of inverting a matrix. If the matrix has some special structure, like being diagonal or lower/upper triangular, the matrix inversion is computationally cheaper than the general case. If an algorithm detects such a structure and adjusts the inversion method accordingly, we call the algorithm adaptive to this regularity.

An adaptive procedure usually consists of two main elements:

1. A *flexible* algorithm: an algorithm that has some tunable parameters and can deliver the "optimal" performance for a vast range of regularities – provided that its parameters are chosen properly.

[2] We have not defined the value function yet, which is formally done in Section 2.1. Readers not familiar with the concept of value functions can read the sentence by replacing "value function" with "target function" in the sense that is commonly used in the regression literature.

[3] More information about the possible difficulties of solving a learning problem and common types of regularities in the supervised learning context is in Appendix A and Section A.2 in particular.

2. A *model selection* algorithm: an algorithm that tunes the parameters of a flexible algorithm.

The usual practice in the RL community is quite different from using adaptive procedures. Typically, an RL user picks a finite pre-defined set of basis functions to represent the value function as a linear combination of the basis. These basis functions are usually chosen a priori by the user and is fixed through learning. This approach, which we call **parametric**, has been thoroughly studied in the RL literature, see e.g., Tsitsiklis and Van Roy [1997]; Sutton et al. [2009]. One important advantage of parametric approaches is that whenever the model is selected properly, such that the true value function can be closely approximated in that model, they show a fast error convergence rate and are often computationally efficient.

Nevertheless, parametric approaches have one serious limitation: if the unknown value function cannot be closely approximated by the parametric model, they show a function approximation error, which may result in poor performance. The usual approach to address this issue is to have a human designer find the right parametric model by fine-tuning the function approximation architecture. For instance, the designer should select the form and the number of basis functions by trial and error. This job is usually difficult, tedious, and against the idea of having a flexible method that can easily work with a large class of functions.

On the other hand, we have flexible **nonparametric** approaches that have much weaker assumptions on the value function. They implicitly or explicitly work with infinite dimensional function spaces and as a result allow us to represent a wide range of value functions. In these approaches, the choice of basis functions themselves may be adaptive and depend on data. Examples of nonparametric methods are K-NN, smoothing kernel estimators, locally linear models, decision trees, growing neural networks, orthogonal series estimates, and regularization-based kernel methods [Györfi et al., 2002; Hastie et al., 2001; Wasserman, 2007; Bishop, 2006].

Nonparametric methods usually have a few tuning parameters.[4] The right choice of these parameters depends on the problem in hand. By changing these parameters data-dependently, one can make them work well for a large range of problems, e.g., for the whole scale of smoothness orders. The downside of nonparametric methods, comparing to parametric ones, is their computational complexity. With the advent of powerful computers and elegant numerical computation algorithms, however, this downside may become less and less of a concern.

One important and powerful class of nonparametric approaches, which has been proven to be an effective tool in statistics and supervised machine learning, is the class of methods that use regularization to control the complexity of a large function space. The main idea is to formulate the learning task as an optimization problem in a large function space where one minimizes the sum of an empirical error and a complexity penalty (the **regularizer**). It is known in the supervised learning/statistics literature that whenever the regularization is selected properly, e.g., by cross-validation or complexity-regularized model selection (also known as structural risk minimization), the resulting procedure automatically adapts to the complexity of the target function, converging almost as fast as if the right model was known beforehand (see e.g., Kohler et al. [2002]; Györfi et al. [2002]). This is the approach we take in this thesis.

1.2 Contributions

The goal of this thesis is to develop flexible regularized value-based algorithms to deal with RL/Planning problems with large state spaces that can be described by a discounted Markov Decision Process. The high-level contributions of this research are:

[4] Note that nonparametric methods are not parameter-free methods.

- Providing regularized algorithms to solve RL/Planning problems based on Approximate Value Iteration (AVI) and Approximate Policy Iteration (API). We formulate these algorithms as regularized optimization problems in large function spaces, and demonstrate how to solve them for the family of **reproducing kernel Hilbert spaces (RKHS)**.
- Introducing a complexity regularization-based algorithm for model selection in RL/ Planning problems.
- Statistical analyzing of the suggested algorithms and providing upper bounds on the performance loss.

Apart from this chapter that motivates the problem, this thesis has five chapters with new contributions (Chapters 3, 4, 5, 6, 7), and three others that supply the reader with the necessary background in sequential decision-making problems (Chapter 2), supervised learning problems (Appendix A), and mathematics (Appendix B).[5] Chapter 8 summarizes the thesis, highlights its limitations such as the accessibility of the state assumption and computational considerations, and suggests several possibilities for future investigations that have been laid by this work. In the rest of this section, we summarize the contributions of each chapter.

Error Propagation for Approximate Policy and Value Iteration (Chapter 3)

This chapter addresses the basic question of how the approximation error/Bellman residual at each iteration of the API/AVI algorithms influences the quality of the resulting policy. The results of this chapter are crucial in the analysis of regularized RL algorithms introduced in Chapters 5 and 6. We quantify the performance loss as the L_p-norm of the approximation error/Bellman residual at each iteration. We also show that the performance loss depends on the expectation of the squared Radon-Nikodym derivative of a certain distribution rather than its supremum – as opposed to what has been suggested by the previous results. Additionally, our results indicate that the contribution of the approximation/Bellman error to the performance loss is more prominent in the later iterations of API/AVI, and the effect of an error term in the earlier iterations decays exponentially fast.[6]

Regularized Least-Squares Regression: Learning from a β-mixing Sequence (Chapter 4)

A main component of our regularized AVI algorithm (Chapter 5) is a regularized least-squares regression estimator. The purpose of this chapter is to prepare for the analysis of RFQI algorithm by providing the rate of convergence of the estimation error in regularized least-squares regression when the data is exponentially β-mixing. The results are proven under the assumption that the metric entropy of the balls in the chosen function space grows at most polynomially. In order to prove our main result, we also derive a relative deviation concentration inequality for β-mixing processes, which might be of independent interest. The other major techniques that we use are the independent-blocks technique and the peeling device. An interesting aspect of our analysis is that in order to obtain fast rates we have to make the block sizes dependent on the layer of peeling. With this approach, up to a logarithmic factor, we recover the optimal minimax rates available for the independent and identically distributed (i.i.d.) case, at least in an asymptotic sense. In particular, our

[5] Even though we have tried to provide a self-contained thesis, there might be places where some background knowledge of statistical machine learning and reinforcement learning/planning is required. The knowledge of reinforcement learning/approximate dynamic programming at the level of Szepesvári [2010], statistical learning theory at the level of Györfi et al. [2002], and machine learning algorithms at the level of Hastie et al. [2001] should suffice.

[6] A version of this chapter has partly been published [Farahmand et al., 2010].

rate asymptotically matches the optimal rate of convergence when the regression function belongs to a Sobolev space.[7]

Regularized Fitted Q-Iteration Algorithm (Chapter 5)

Regularized Fitted Q-Iteration (**RFQI**) is a novel nonparametric AVI algorithm to solve RL/Planning problems with large state spaces. RFQI uses regularized least-squares regression to approximately perform a single step of the value iteration.

To analyze the statistical properties of RFQI and provide an error upper bound on the performance loss of the resulting policy, we provide an upper bound on the L_2-norm of the fitting error at each iteration. To provide such a bound, we need to not only analyze the convergence behavior of the regularized regression algorithm (Chapter 4), but also consider the effect of previous iterations on the current one. This effect is in the form of changing both the smoothness of the target function and the function approximation error. Afterwards, the result of Chapter 3 can be applied to provide an upper bound on the performance loss of the resulting policy. The main result, Theorem 5.8, provides a performance loss upper bound of the resulting policy and shows its dependence on the number of samples, the choice of the function space, and some intrinsic properties of the underlying MDP. The result indicates that by the appropriate choice of the function spaces and the regularization coefficients, achieving rates as fast as the optimal minimax convergence rate for certain classes of RKHS is possible.[8]

Regularized Policy Iteration Algorithm (Chapter 6)

We introduce two nonparametric Approximate Policy Iteration algorithms, namely **REG-LSPI** and **REG-BRM**, to solve reinforcement learning and planning problems with large state spaces. Our algorithms are built on the regularized extensions of the Least-Squares Temporal Difference (LSTD) learning and the Bellman Residual Minimization (BRM) procedures for policy evaluation. We derive efficient implementations of our methods when the function space is a reproducing kernel Hilbert space. We also analyze the statistical properties of REG-LSPI and provide an upper bound on the policy evaluation error and the performance loss of the resulting policy. We show how this error depends on the number of samples, the capacity of the function space, and some intrinsic properties of the underlying MDP. The dependence of the policy evaluation bound on the number of samples is minimax optimal.[9]

Model Selection in Reinforcement Learning (Chapter 7)

We consider the problem of model selection in the batch (offline, non-interactive) reinforcement learning setting when the goal is to find an action-value function with the smallest Bellman error among a countable set of candidates functions. We propose a complexity regularization-based model selection algorithm, BErMin, and prove that it enjoys an oracle-like property: the estimator's error differs from that of an oracle, who selects the candidate with the minimum Bellman error, by only a constant factor and a small remainder term that vanishes at a parametric rate as the number of samples increases. As an application, we consider a problem when the true action-value function belongs to an unknown member of a nested sequence of function spaces. We show that under some additional technical conditions BErMin leads to a procedure whose rate of convergence, up to a constant factor, matches that of an oracle who knows to which of the nested function spaces the true action-value function belongs, i.e., the procedure achieves adaptivity.[10]

[7] A version of this chapter has been published [Farahmand and Szepesvári, 2012].
[8] Some versions of this chapter have partly been published [Farahmand et al., 2008, 2009a,c].
[9] A version of this chapter has partly been published [Farahmand et al., 2009b].
[10] A version of this chapter has been published [Farahmand and Szepesvári, 2011].

1.3 Credits

I acknowledge the great help and contributions of Csaba Szepesvári, Mohammad Ghavamzadeh, Shie Mannor, and Rémi Munos. Although I have been directly involved in most parts of this research program, some parts have not been studied, proven, or written by me, or to them I had only minor contributions. For the sake of completeness, however, I include them in this thesis. These results are as follows.

- The matrix form of Theorem 6.3 is derived by Mohammad Ghavamzadeh and Csaba Szepesvári. I was contributing to discussions about the new representer theorem, but I have not derived the formula myself.
- Theorem 7.3 is stated and proven mostly by Csaba Szepesvári.
- Figure 4.1 is drawn by Csaba Szepesvári.

Chapter 2

Sequential Decision-Making Problems

This chapter begins by providing the necessary background on sequential decision-making problems. We define the mathematical framework of Markov Decision Processes (MDP) in Section 2.1, and afterwards we introduce Reinforcement Learning (RL) and Dynamic Programming (DP)-based planning problems in Section 2.2. These two problems are very similar with the exception that they describe situations with different prior knowledge about the problem in hand. We describe the value-based approach to solve RL/Planning problems in Section 2.3 and briefly review methods such as Value Iteration and Policy Iteration algorithms. In Section 2.4, we explain two common ways to measure the performance of RL/Planning algorithms. Finally in Section 2.5, we discuss difficulties of solving RL/Planning problems in large state spaces where one has to use function approximation. There we categorize different algorithms according to their modeling assumption (parametric vs. nonparametric) and their statistical convergence behavior.

Several textbooks and monographs on RL and Planning provide comprehensive reviews of these problems and associated algorithms. Sutton and Barto [1998] is an introductory-level textbook that covers both RL and Planning, with more emphasis on the learning aspects. Sutton and Barto consider both discrete and continuous state spaces. Bertsekas and Tsitsiklis [1996] is a more advanced textbook on RL and Planning. Bertsekas and Shreve [1978] is an advanced monograph on Planning that provides a treatment on general state spaces, both finite and infinite, but does not cover learning/estimation aspect of the problem. Bertsekas [2010] is a work in progress chapter that covers recent advances of "Approximate Dynamic Programming" and has similar style as Bertsekas and Tsitsiklis [1996]. And finally, Buşoniu et al. [2010a] and Szepesvári [2010] are two new monographs that cover recent developments in the RL/Planning literature.

2.1 Definitions

Probability Space

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

We write $\nu_1 \ll \nu_2$ if $\nu_2(A) = 0$ implies that $\nu_1(A) = 0$ as well. For two σ-finite measures ν_1 and ν_2 on some measurable space (Ω, σ_Ω), ν_1 is *absolutely continuous* w.r.t. ν_2 if there is a non-negative measurable function $f : \Omega \to \mathbb{R}$ such that $\mu_1(A) = \int f d\nu_2$ for all $A \in \sigma_\Omega$. It is known that ν_1 is absolutely continuous w.r.t. ν_2 if and only if $\nu_1 \ll \nu_2$. We write $\frac{d\nu_1}{d\nu_2} = f$

and call it the *Radon-Nikodym* derivative of ν_1 w.r.t. ν_2 [Rosenthal, 2006, Chapter 12].

Markov Decision Process

Definition 2.1. *A* finite-action discounted MDP *is a 4-tuple $(\mathcal{X}, \mathcal{A}, P, \gamma)$, where $\mathcal{X}$ is a measurable state space, $\mathcal{A} = \{a_1, a_2, \ldots, a_{|\mathcal{A}|}\}$ is the finite set of $|\mathcal{A}|$ actions, $P : \mathcal{X} \times \mathcal{A} \to \mathcal{M}(\mathbb{R} \times \mathcal{X})$ is the reward-transition probability kernel with domain $\mathcal{X} \times \mathcal{A}$ and $0 \leq \gamma < 1$ is the 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 $P(\cdot|x, a) = P_{x,a}(\cdot) = \int_{\mathbb{R}} P(dr, \cdot|x, a)$ (transition probability kernel) and $\mathcal{R}(\cdot|x, a) = \int_{\mathcal{X}} P(\cdot, dy|x, a)$ (reward distribution).*

MDPs encode the temporal evolution of a discrete-time stochastic process controlled by an *agent*. The dynamical system starts at time $t = 1$ with random initial state $X_1 \sim P_1$ where " $\sim$ " denotes that X_1 is drawn from distribution P_1. At time t, action $A_t \in \mathcal{A}$ is selected by the agent controlling the process. As a result the pair (R_t, X_{t+1}) is drawn from $P(\cdot, \cdot|X_t, A_t)$, i.e., $(R_t, X_{t+1}) \sim P(\cdot, \cdot|X_t, A_t)$. Here, R_t is the reward that the agent receives at time t and X_{t+1} is the state at time $t + 1$. This procedure continues and leads to a random *trajectory* $\xi = (X_1, A_1, R_1, X_2, A_2, R_2, \cdots)$. We denote the space of all possible trajectories as Ξ.

This definition of MDP is quite general. If $\mathcal{X}$ is a finite state space, the result is called a finite MDP. The state space $\mathcal{X}$ can be more general. If we consider a measurable subset of $\mathbb{R}^d$ ($\mathcal{X} \subseteq \mathbb{R}^d$), such as $(0, 1)^d$, we get the so-called continuous state-space MDPs. In this thesis, we often talk about measurable subsets of $\mathbb{R}^d$, but one can think of other state spaces too, e.g., the binary lattices $\{0, 1\}^d$, the space of graphs, the space of strings, the space of distributions, etc.

Policy

Definition 2.2 (Definition 8.2 and 9.2 of Bertsekas and Shreve [1978])**.** *A* ***policy*** *is a sequence $\bar{\pi} = \{\pi_1, \pi_2, \ldots\}$ such that for each t,*

$$\pi_t(a_t|X_1, A_1, X_2, A_2, \ldots, X_{t-1}, A_{t-1}, X_t)$$

is a universally measurable stochastic kernel on $\mathcal{A}$ given $\underbrace{\mathcal{X} \times \mathcal{A} \times \cdots \times \mathcal{X} \times \mathcal{A} \times \mathcal{X}}_{2t-1 \text{ elements}}$ satisfying

$$\pi_t(\mathcal{A}|X_1, A_1, X_2, A_2, \ldots, X_{t-1}, A_{t-1}, X_t) = 1$$

for every $(X_1, A_1, X_2, A_2, \ldots, X_{t-1}, A_{t-1}, X_t)$. If π_t is parametrized only by X_t, $\bar{\pi}$ is a Markov *policy. If for each t and $(X_1, A_1, X_2, A_2, \ldots, X_{t-1}, A_{t-1}, X_t)$, the policy π_t assigns mass one to a single point in $\mathcal{A}$, $\bar{\pi}$ is called a* deterministic (nonrandomized) *policy; if it assigns a distribution over $\mathcal{A}$, it is called* stochastic *or* randomized *policy. If $\bar{\pi}$ is a Markov policy in the form of $\bar{\pi} = (\pi, \pi, \ldots)$, it is called a* stationary *policy.*

We define the following terminology and notations in order to simplify our exposition.

Definition 2.3. *We say that an agent is "following" a Markov stationary policy π whenever A_t is selected according to the policy $\pi(\cdot|X_t)$, i.e., $A_t \sim \pi(\cdot|X_t)$. The policy π induces two transition probability kernels $P^\pi : \mathcal{X} \to \mathcal{M}(\mathcal{X})$ and $P^\pi : \mathcal{X} \times \mathcal{A} \to \mathcal{M}(\mathcal{X} \times \mathcal{A})$. For a measurable subset A of $\mathcal{X}$ and a measurable subset B of $\mathcal{X} \times \mathcal{A}$, denote*

$$(P^\pi)(A|x) \triangleq \int_{\mathcal{X}} P(dy|x, \pi(x)) \mathbb{I}_{\{y \in A\}},$$
$$(P^\pi)(B|x, a) \triangleq \int_{\mathcal{X}} P(dy|x, a) \mathbb{I}_{\{(y, \pi(y)) \in B\}}.$$

The m-step transition probability kernels $(P^\pi)^m : \mathcal{X} \to \mathcal{M}(\mathcal{X})$ and $(P^\pi)^m : \mathcal{X} \times \mathcal{A} \to \mathcal{M}(\mathcal{X} \times \mathcal{A})$ for $m = 2, 3, \cdots$ are inductively defined as

$$(P^\pi)^m(A|x) \triangleq \int_{\mathcal{X}} P(dy|x, \pi(x))(P^\pi)^{m-1}(A|y),$$
$$(P^\pi)^m(B|x, a) \triangleq \int_{\mathcal{X}} P(dy|x, a)(P^\pi)^{m-1}(B|y, \pi(y)).$$

The difference between the transition probability kernels $P^\pi : \mathcal{X} \to \mathcal{M}(\mathcal{X})$ and $P^\pi : \mathcal{X} \times \mathcal{A} \to \mathcal{M}(\mathcal{X} \times \mathcal{A})$ is in the way the policy affects the action selection: in the former, the action of the first step is chosen according to the policy, while in the latter the first action is pre-chosen and the policy chooses the action in the second step.

Definition 2.4. *Given probability transition kernels $P : \mathcal{X} \to \mathcal{M}(\mathcal{X})$ and $P : \mathcal{X} \times \mathcal{A} \to \mathcal{M}(\mathcal{X} \times \mathcal{A})$, define the right-linear operators $P\cdot : B(\mathcal{X}) \to B(\mathcal{X})$ and $P\cdot : B(\mathcal{X} \times \mathcal{A}) \to B(\mathcal{X} \times \mathcal{A})$ by*

$$(PV)(x) \triangleq \int_{\mathcal{X}} P(dy|x)V(y),$$
$$(PQ)(x, a) \triangleq \int_{\mathcal{X} \times \mathcal{A}} P(dy, da'|x, a)Q(y, a').$$

For a probability measure $\rho \in \mathcal{M}(\mathcal{X})$ and a measurable subset A of $\mathcal{X}$, define the left-linear operator $\cdot P : \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{X})$ by

$$(\rho P)(A) = \int \rho(dx)P(dy|x)\mathbb{I}_{\{y \in A\}}.$$

In words, ρP represents the distribution over states when the initial state distribution is ρ and we follow P for a single step.
Similarly, for a probability measure $\rho \in \mathcal{M}(\mathcal{X} \times \mathcal{A})$ and a measurable subset B of $\mathcal{X} \times \mathcal{A}$, define the left-linear operator $\cdot P : \mathcal{M}(\mathcal{X} \times \mathcal{A}) \to \mathcal{M}(\mathcal{X} \times \mathcal{A})$ by

$$(\rho P)(B) = \int \rho(dx, da)P(dy, da'|x, a)\mathbb{I}_{\{(y,a') \in B\}}.$$

A typical choice of P is $(P^\pi)^m$ for $m \geq 1$.

Under certain conditions, it can be shown that a deterministic Markov stationary policy is all we should care for, e.g., see Proposition 4.3 of Bertsekas and Shreve [1978]. From now on, whenever we use term "policy", we are referring to a *deterministic Markov stationary policy* and we denote it by π (instead of $\bar{\pi}$) – unless it is stated otherwise.

Planning and Reinforcement Learning as a Variational Problem

In a non-orthodox viewpoint, reinforcement learning and planning problems can be seen as maximizing a functional of the reward distribution $\mathcal{R}(\cdot|x, a)$. Let $G : \Xi \to \mathbb{R}$ be the *return* function that is defined by the designer of the sequential decision-making problem. Let $\xi(x)$ be a trajectory starting from x, and denote $P^\pi_{\xi(x)}$ as the probability measure induced by the policy π on the space of all trajectories starting from x. Define the following functional:

$$\mathcal{J}(x; \pi, P, G) \triangleq \int_{\Xi} G(\xi)dP^\pi_{\xi(x)}(\xi).$$

The goal of planning and reinforcement learning is to find a policy π^* that maximizes this functional (if there exists any), i.e.,

$$\mathcal{J}(\cdot; \pi^*, P, G) = \sup_\pi \mathcal{J}(\cdot; \pi, P, G).$$

We call π^* an *optimal* policy.

Discounted MDPs

One specific type of functionals that deserves special attention is the *discounted* reward functional. Under this functional the importance of the future reward is less than the imminent one. In addition to its suitability to model some classes of sequential decision-making problems, the resulting mathematics is often easier to analyze.

For a given policy π, let $\xi(x) = (X_1 = x, \pi(X_1), R_1, X_2, \pi(X_2), R_2, \cdots)$ be the sequence induced by following policy π from the initial state x. Define the γ-discounted return function as $G_\gamma(\xi) = \sum_{t=1}^{\infty} \gamma^{t-1} R_t$. The discounted reward functional is defined as[1]

$$\mathcal{J}_\gamma(x; \pi, P, G) \triangleq \int_\Xi G_\gamma(\xi) dP^\pi_{\xi(x)}(\xi) = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} R_t\right]$$

Bertsekas and Shreve [1978, Proposition 7.45]) guarantees the well-definedness of this expectation. Discounted MDPs will be the focus of our further developments.

Value Functions

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

Definition 2.5 (Value Functions)**.** *The value function V^π and the action-value function Q^π for a policy π are defined as follows: Let $(R_t; t \geq 1)$ be the sequence of rewards when the process is started from a state X_1 (or (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 follows the policy π for $t \geq 1$ (or $t \geq 2$ for the action-value function). Then,*

$$\begin{aligned} V^\pi(x) &\triangleq \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} R_t | X_1 = x\right], \\ Q^\pi(x, a) &\triangleq \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} R_t | X_1 = x, A_1 = a\right]. \end{aligned}$$

In words, the value function V^π evaluated at state x is the expected discounted return of following the policy π from state x. The action-value function evaluated at (x, a) is the expected discounted return when the agent starts at state x, takes action a, and then follows policy π.

For a discounted MDP, we define the *optimal value function* by

$$V^*(x) \triangleq \sup_\pi V^\pi(x), \qquad \text{(for all } x \in \mathcal{X})$$

and similarly the *optimal action-value function* is defined as

$$Q^*(x, a) \triangleq \sup_\pi Q^\pi(x, a). \qquad \text{(for all } (x, a) \in \mathcal{X} \times \mathcal{A})$$

We say that a deterministic policy π is *greedy* w.r.t. an action-value function Q (or a value function V) and write $\pi = \hat{\pi}(\cdot; Q)$ (or $\pi = \hat{\pi}(\cdot; V)$), if for all $x \in \mathcal{X}$,

$$\begin{aligned} \pi(x) &= \arg\max_{a \in \mathcal{A}} Q(x, a), && \text{(action-value function)} \\ \pi(x) &= \operatorname*{argmax}_{a \in \mathcal{A}} \int P(dy|x, a)[r(x, a) + \gamma V(y)]. && \text{(value function)} \end{aligned}$$

[1] The reward functional is sometimes defined as $\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t\right]$ too. For notational convenience in our later developments, we have chosen the current indexing.

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. Q^* (or V^*) is an optimal policy. Hence, knowing Q^* is sufficient for behaving optimally [Bertsekas and Shreve, 1978, Proposition 4.3].

Define the immediate expected reward function

$$r(x,a) = \int r\, \mathcal{R}(dr|x,a).$$

It is easy to see that for any policy π, if the absolute value of the immediate expected reward $r^\pi(x) = r(x,\pi(x))$ (or $r^\pi(x) = \sum_{a\in\mathcal{A}} r(x,a)\pi(a|x)$ for stochastic policies) is uniformly bounded by $R_{\max}$, the functions V^π and Q^π are bounded by $V_{\max} = Q_{\max} = R_{\max}/(1-\gamma)$, independent of the choice of π. Moreover, if for all policies π the value of $R_{\max}$ is a uniform upper bound for r^π, V^* and Q^* are also upper bounded by $V_{\max}$.

Bellman Operators

Bellman [optimality] operators provide a useful way to describe and analyze the properties of MDPs. They are particularly important because their fixed points are [optimal] value functions. Proposition 4.2 of Bertsekas and Shreve [1978] shows the optimality of the fixed point of the Bellman optimality operators. Moreover, it shows the uniqueness of the fixed point for both the Bellman and the Bellman optimality operators.

Definition 2.6 (Bellman Operators)**.** *The Bellman operators $T^\pi : B(\mathcal{X}) \to B(\mathcal{X})$ (for the value function V) and $T^\pi : B(\mathcal{X}\times\mathcal{A}) \to B(\mathcal{X}\times\mathcal{A})$ (for the action-value function Q) for the policy π are defined as*

$$\begin{aligned}(T^\pi V)(x) &\triangleq r^\pi(x) + \gamma \int_{\mathcal{X}} V(y) P(dy|x,\pi(x)),\\ (T^\pi Q)(x,a) &\triangleq r(x,a) + \gamma \int_{\mathcal{X}} Q(y,\pi(y)) P(dy|x,a).\end{aligned} \tag{2.1}$$

The fixed point of this operator is the [action-]value function of the policy π, i.e., $T^\pi Q^\pi = Q^\pi$ and $T^\pi V^\pi = V^\pi$ (Proposition 4.2(b) of Bertsekas and Shreve [1978]).

Definition 2.7 (Bellman Optimality Operators)**.** *The Bellman optimality operators $T^* : B(\mathcal{X}) \to B(\mathcal{X})$ and $T^* : B(\mathcal{X}\times\mathcal{A}) \to B(\mathcal{X}\times\mathcal{A})$ are defined as*

$$\begin{aligned}(T^* V)(x) &\triangleq \max_a \left\{ r(x,a) + \gamma \int_{\mathcal{X}} V(y) P(dy|x,a) \right\},\\ (T^* Q)(x,a) &\triangleq r(x,a) + \gamma \int_{\mathcal{X}} \max_{a'} Q(y,a') P(dy|x,a).\end{aligned} \tag{2.2}$$

These operators enjoy a fixed-point property similar to that of the Bellman operators: $T^*Q^* = Q^*$ and $T^*V^* = V^*$ [Bertsekas and Shreve, 1978, Proposition 4.2(a)].

Proposition 4.3 of Bertsekas and Shreve [1978] implies that the optimal value function can be attained by a deterministic Markov stationary policy if the action set is finite. This result also holds for infinite action sets when certain compactness conditions are satisfied [Bertsekas and Shreve, 1978, Proposition 4.4]. As we only focus on the MDPs with finite action sets, we do not report the detail of these conditions.

Norms and Function Spaces

We use $\mathcal{F} : \mathcal{X} \to \mathbb{R}$ to denote a subset of measurable functions. The exact specification of this space will be clear from the context. We usually denote $\mathcal{F}$ as the space of value functions.

For a probability measure $\nu_{\mathcal{X}} \in \mathcal{M}(\mathcal{X})$, and a measurable function $V \in \mathcal{F}$, we define the $L_p(\nu_{\mathcal{X}})$-norm of V as

$$\|V\|_{p,\nu_{\mathcal{X}}}^p \triangleq \int_{\mathcal{X}} |V(x)|^p d\nu_{\mathcal{X}}(x). \tag{2.3}$$

The $L_\infty(\mathcal{X})$-norm is defined as $\|V\|_\infty \triangleq \sup_{x\in\mathcal{X}} |V(x)|$.

We define $\mathcal{F}^{|\mathcal{A}|} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^{|\mathcal{A}|}$ as a subset of vector-valued measurable functions with the following identification:

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

We use $Q_j(x) = Q(x, j)\,(j = 1, \dots, |\mathcal{A}|)$ to refer to the j^{th} component of $Q \in \mathcal{F}^{|\mathcal{A}|}$. We often denote $\mathcal{F}^{|\mathcal{A}|}$ as the space of action-value functions. For $\nu \in \mathcal{M}(\mathcal{X} \times \mathcal{A})$ and $Q \in \mathcal{F}^{|\mathcal{A}|}$, we define $\|\cdot\|_{p,\nu}$ by generalizing (2.3) to $\mathcal{F}^{|\mathcal{A}|}$ as follows

$$\|Q\|_{p,\nu}^p \triangleq \int_{\mathcal{X}\times\mathcal{A}} |Q(x, a)|^p d\nu(x, a). \tag{2.4}$$

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

$$\|f\|_{p,z_{1:n}}^p = \|f\|_{p,\mathcal{D}_n}^p \triangleq \frac{1}{n} \sum_{i=1}^{n} |f(z_i)|^p. \tag{2.5}$$

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

Note that we sometimes use the shorthand notation of $\nu|Q|^p = \|Q\|_{p,\nu}^p$ (similar for $\nu_{\mathcal{X}}$ and other probability distributions). In this thesis most, but not all, results are stated for $p = 1$ or $p = 2$. The symbols $\|\cdot\|_\nu$ and $\|\cdot\|_n$ refer to an L_2-norm.

Finally, define the projection operator $\Pi_{\nu,\mathcal{F}^{|\mathcal{A}|}} : B(\mathcal{X} \times \mathcal{A}) \to B(\mathcal{X} \times \mathcal{A})$ as $\Pi_{\nu,\mathcal{F}} Q \triangleq \operatorname{argmin}_{Q' \in \mathcal{F}^{|\mathcal{A}|}} \|Q' - Q\|_\nu^2$ for $Q \in B(\mathcal{X} \times \mathcal{A})$. The definition of $\Pi_{\nu_{\mathcal{X}},\mathcal{F}} : B(\mathcal{X}) \to B(\mathcal{X})$ is similar. If the measures $\nu_{\mathcal{X}}$ or ν are clear from the context, we may simply write $\Pi_{\mathcal{F}}$ and $\Pi_{\mathcal{F}^{|\mathcal{A}|}}$ instead.

2.2 Reinforcement Learning and Planning

Reinforcement Learning and Planning are two similar types of sequential decision-making problems with the common goal of finding a policy π that has the performance equal or close to that of the optimal policy π^*. The difference between reinforcement learning and planning problems, as we will discuss shortly, is in our prior knowledge about the problem and the way we interact with it. In this thesis, we only focus on problems that can be modeled by an MDP.

In *Planning* the transition probability kernel $P(\cdot|x, a)$ and the reward distribution $\mathcal{R}(\cdot|x, a)$ of the MDP is known. On the other hand, in *Reinforcement Learning* either P or $\mathcal{R}$ or even both are not directly accessible, but one interacts with the MDP by selecting action A_t at state X_t, and getting a reward $R_t \sim \mathcal{R}(\cdot|X_t, A_t)$ and going to the next state X_{t+1} according to the transition probability kernel. The result is a trajectory $\xi = (X_1, A_1, R_1, X_2, A_2, R_2, \cdots)$. This mode of interaction is usually described by an *agent-environment* metaphor in the RL community [Sutton and Barto, 1998].

There are some middle ground scenarios as well. Sometimes we do have the luxury of knowing $P(\cdot, \cdot|x, a)$ but cannot compute functionals involving it, such as $T^\pi Q$, because of

the large cardinality of $\mathcal{X}$. Another situation is when we do not have access to $P(\cdot, \cdot|x, a)$, but have access to a flexible data generator that gets any $(x, a) \in \mathcal{X} \times \mathcal{A}$ as input and returns $(R, X') \sim P(\cdot, \cdot|x, a)$. Finding a good policy in these scenarios is called the problem of *Approximate Planning*.

Several approaches to solve RL/Planning problems exist. Based on the type of explicit representation that these approaches maintain, one may categorize them into two major classes:

- Value Space Search
- Policy Space Search

Value-based approaches maintain an estimate $\hat{Q}$ (or $\hat{V}$) of the optimal value function Q^* (or V^*). The premise of value-based approaches is that by finding an accurate enough estimate $\hat{Q}$ of the optimal action-value function $\hat{Q}$, the greedy policy $\hat{\pi}(\cdot; \hat{Q})$ will be close to the optimal policy in some well-defined sense. On the other hand, the direct policy search approaches explicitly represent the policy function and directly perform the search in the policy space. The search may be guided by the gradient information [Baxter and Bartlett, 2001; Kakade, 2001; Ghavamzadeh and Engel, 2007b] or be in the same spirit as evolutionary algorithms [Moriarty et al., 1999; Heidrich-Meisner and Igel, 2009]. Moreover, there are hybrid methods that explicitly represent both value and policy functions [Konda and Tsitsiklis, 2001; Peters et al., 2003; Ghavamzadeh and Engel, 2007a]. In this work, we only focus on the value-based approaches.

2.2.1 Online vs. Offline Samples; Batch vs. Incremental Processing

An important aspect of any method that solves RL/Planning problems is the way that data are collected and processed by the algorithm. The data collection setting can be categorized as *online* or *offline* and the data processing method can be categorized as *batch* or *incremental*.

Online vs. Offline Sampling

The online sampling setting is when the agent chooses the action sequence $A_t \sim \pi_t$ and directly influences how the data stream $\xi = (X_1, A_1, R_1, \dots)$ is generated. The offline setting, on the other hand, is when the agent does not have control over how the data are generated; the agent is, rather, provided with a data set[2]

$$\mathcal{D}_n = \{(X_1, A_1, R_1, X'_1), \dots, (X_n, A_n, R_n, X'_n)\}, \tag{2.6}$$

where $(R_i, X'_i) \sim P(\cdot, \cdot|X_i, A_i)$, $A_i \sim \pi_b(\cdot|X_i)$, and $X_i \sim \nu_{\mathcal{X}}$ $(i = 1, \dots, n)$, with $\nu_{\mathcal{X}}$ as the fixed distribution over the states. The policy π_b is the data-generating policy and is commonly known as the *"behavior"* policy. The behavior policy is usually a stochastic one, and might be unknown to the agent. We shall denote by ν the common distribution underlying (X_i, A_i). Samples X_i and X_{i+1} may be sampled independently (common in the Planning scenario), or may be coupled through $X'_i = X_{i+1}$ (common in the RL scenario). In the latter case the data belong to a single trajectory. Under either of these assumptions we say that the data $\mathcal{D}_n$ meet the *standard offline sampling assumption*.

Batch vs. Incremental Processing

An algorithm can be batch or incremental. A batch algorithm processes the whole data set $\mathcal{D}_n$ and can freely access any element of the data set at any time. An incremental algorithm, however, continues to learn whenever a new data sample is available. The computation does not directly depend on the whole data set $\mathcal{D}_n$, but only on the recent data sample (X_n, A_n, R_n, X'_n). Of course, the boundary between a batch algorithm and an incremental one is not vividly clear. One may say an incremental algorithm is a special case of the batch algorithms when the algorithm processes data in a specific temporal ordering.

The question of which of these settings is more natural depends on the problem in hand. If all available is a collection of data $\mathcal{D}_n$, and interacting with the MDP is impossible, we are

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

inevitably in the offline setting. In this case as the batch algorithms are usually more data efficient they are the preferred choice for data processing – unless the computation time is limited. On the other hand, if we have direct access to the environment, either by knowing the model of the MDP or accessing its generative model (as is common in planning) or when the agent is actually situated in the environment, the data sampling scenario is indeed online and both batch and incremental algorithms may be used. In this work, we focus on the batch algorithms that assume $\mathcal{D}_n$ meets the standard offline sampling assumption. The assumption that the states $\{X_i\}$ are identically distributed and that a stationary policy π_b is used to generate the data can be relaxed but would complicate the analysis. Hence for simplicity, we stick to the above assumptions in the rest of this work.

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

Definition 2.8 (Empirical Bellman Operators)**.** *Let $\mathcal{D}_n$ be a dataset as (2.6). Define the ordered multiset $S_n = \{(X_1, A_1), \dots, (X_n, A_n)\}$. For a given fixed policy π, the empirical Bellman operator $\hat{T}^\pi : \mathbb{R}^{S_n} \to \mathbb{R}^n$ is defined as*

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

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

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

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 value function Q at the selected points.

The following proposition, which follows immediately from the definitions, shows that the empirical Bellman operators provide an unbiased estimate of the Bellman operators (Note that $\hat{T}^\pi$ and $\hat{T}^*$ depend on the data, and hence they are random. The dependence is suppressed to simplify the notation).

Proposition 2.1. *For any fixed, bounded measurable, deterministic function $Q : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, policy π and index $1 \le i \le n$, it holds that*

$$\mathbb{E}\left[\hat{T}^\pi Q(X_i, A_i) \,\middle|\, X_i, A_i\right] = T^\pi Q(X_i, A_i),$$
$$\mathbb{E}\left[\hat{T}^* Q(X_i, A_i) \,\middle|\, X_i, A_i\right] = T^* Q(X_i, A_i),$$

2.3 Value-based Approaches for Reinforcement Learning and Planning

In the value-based approaches for solving RL/Planning problems, we aim to find the [approximate] fixed point of the Bellman operator $Q^\pi = T^\pi Q^\pi$ (or $V^\pi = T^\pi V^\pi$) for the so-called *policy evaluation* problem or the Bellman optimality operator $Q^* = T^* Q^*$ (or $V^* = T^* V^*$). To find a close to optimal value function, we are facing the following challenges:

1. How to represent an action-value function Q?
2. Given Q, how to evaluate $T^\pi Q$ or $T^* Q$?
3. How to find the fixed point of T^π or T^* operators?

The first problem is easy when $\mathcal{X} \times \mathcal{A}$ is a small finite space, so Q can be represented by a finite number of real values. When it is not, we must approximate Q with a simpler and easier to compute function called *approximant*. The process of approximating a function with an easier to compute function is called *Function Approximation (FA)*, which different

aspects of it are studied in the approximation theory [Devore, 1998] and the statistical learning theory [Györfi et al., 2002].

The second challenge is to evaluate $T^\pi Q$ or $T^* Q$ given Q. Here one requires to calculate integrals of (2.1) and (2.2). Except special cases, such as in the Linear Quadratic Regulation (LQR) problem [Burl, 1998], this is intractable for large state spaces, even if $P(\cdot, \cdot|x, a)$ is known. A reasonable way to evaluate $T^\pi Q$ or $T^* Q$, for both RL/Planning scenarios, is to approximately estimate them by random sampling from $P(\cdot, \cdot|x, a)$.

The third challenge is to find the fixed point of the Bellman operators. There are several approaches to solve this problem. In the following subsection, we briefly mention some important families of methods for finding the fixed point of the Bellman [optimality] operator.

2.3.1 Generic Solution Methods

Linear System of Equations and Linear Programming

For an MDP with a finite number of states and actions, policy evaluation problem is equivalent to solving the finite linear system of equations described by $Q = T^\pi Q$. To find the fixed point of the Bellman *optimality* operator, however, one has to solve a non-differentiable nonlinear optimization problem. The equation $Q^* = T^* Q^*$, however, can be cast as a Linear Programming (LP) problem. Exact solution of either a linear systems of equations or LP is feasible only for small MDPs. Approaches based on the approximate LP have been investigated in the literature (see e.g., Schuurmans and Patrascu [2001]; de Farias and Van Roy [2003]; Petrik et al. [2010]), but we do not study them in this work.

One popular approach to find the fixed point of the Bellman operator is to benefit from its contraction or monotonicity properties. Briefly speaking, these properties imply that one may find the fixed point of the Bellman operator by an iterative procedure such as *Value Iteration (VI)* or *Policy Iteration (PI)* (see Bertsekas and Shreve [1978] and Szepesvári [1997b] for details of the conditions that guarantee these methods to work).

Value Iteration

Value Iteration is an iterative method to find the fixed point of the Bellman [optimality] operator by benefiting from the *contraction* property of these operators. The algorithm starts from an initial value function Q_0 (or likewise V_0), and iteratively applies T^* (or T^π for the policy evaluation problem) to the previous estimate:

$$Q_{k+1} = T^* Q_k.$$

It is known that $\lim_{k\to\infty} (T^*)^k Q_0 = Q^*$ and $\lim_{k\to\infty} (T^\pi)^k Q_0 = Q^\pi$ for every Q_0 (see Proposition 2.6 of Bertsekas and Tsitsiklis [1996] for the result for finite MDPs; Proposition 4.2(c) of Bertsekas and Shreve [1978] for a more general result). For discrete state and action spaces, value iteration may also be performed asynchronously. If we define $TQ|_{\mathcal{X}'\times\mathcal{A}'}$ as the operator TQ restricted to $\mathcal{X}' \times \mathcal{A}' \subset \mathcal{X} \times \mathcal{A}$, we still have the same convergence guarantee provided that all components are chosen infinitely often (Proposition 2.3 of Bertsekas and Tsitsiklis [1996]).

Sometimes, especially when the state-action space is large, this procedure can only be performed approximately, i.e.,

$$Q_{k+1} \approx T^* Q_k.$$

In this case, we call the procedure the *Approximate Value Iteration (AVI)*. Analyzing AVI to determine how the approximation error influences the resulting policy is the topic of Chapters 3 and 5. Some examples of AVI are tree-based Fitted Q-Iteration of Ernst et al. [2005], multi-layer perceptron-based Fitted Q-Iteration of Riedmiller [2005], and Fitted Q-Iteration for continuous action spaces of Antos et al. [2008a]. See the work of Munos and Szepesvári [2008] for more information on AVI.

Policy Iteration

Policy Iteration is another iterative method to find the fixed point of the Bellman *optimality* operator. It starts from a policy π_0, and then *evaluates* it to find Q^{π_0}, i.e., finding a Q_0 that satisfies $T^{\pi_0} Q^{\pi_0} = Q^{\pi_0}$. This is called the *Policy Evaluation* step. Following that, the policy iteration algorithm obtains the greedy policy w.r.t. the most recent value function

$\pi_1 = \hat{\pi}(\cdot; Q^{\pi_0})$. This is called the *Policy Improvement* step. The policy iteration algorithm continues by evaluating the newly obtained policy π_1, and repeating the whole process again, to generate a sequence of policies and their corresponding action-value functions

$$Q^{\pi_0} \to \pi_1 \to Q^{\pi_1} \to \pi_2 \to \dots .$$

Bertsekas and Shreve [1978, Proposition 4.8] shows that for finite state/action MDPs, whenever the policy evaluation step of PI is done precisely, PI yields the optimal policy after a finite number of iterations. Similarly, Bertsekas and Shreve [1978, Proposition 4.9] indicates that a slightly modified policy iteration algorithm, where there is a possibility of having certain amount of error in policy evaluation step, terminates in a finite number of iterations and the value of the resulting policy is close to the optimal one. For more information on the computational complexity of the Value/Policy Iteration algorithms, refer to Ye [2010].

For the policy evaluation step of PI, one requires to solve $T^{\pi_k} Q^{\pi_k} = Q^{\pi_k}$ for a given π_k. For small problems, one may directly solve the system of linear equations as described earlier. For large problems, which is our main interest, one can only approximately solve the policy evaluation step, that is

$$Q_k \approx T^{\pi_k} Q_k$$

We call this scenario the *Approximate Policy Iteration (API)*.

The policy evaluation step of API can be performed in various ways. One possibility is to use AVI to find the fixed point of T^{π_k} operator. Two other important methods, which are the focus of our discussion, are the *Least-Squares Temporal Difference (LSTD)* [Bradtke and Barto, 1996; Lagoudakis and Parr, 2003] and the *Bellman Residual Minimization (BRM)* (Antos et al. [2008b]; Maillard et al. [2010]). When one uses LSTD in the policy iteration algorithm, the resulting method is called the *Least Squares Policy Improvement (LSPI)* [Lagoudakis and Parr, 2003]. In Chapter 3 we analyze the error propagation aspect of API algorithm, and in Chapter 6 we provide and study a new regularization-based formulation of LSTD and BRM that handles problems with large state spaces.

API is a popular approach in the RL literature. Other than the work of Lagoudakis and Parr [2003]; Bradtke and Barto [1996]; Maillard et al. [2010], we would like to mention the work of Kolter and Ng [2009] that formulates an l_1-regularization extension of LSTD, Xu et al. [2007] and Jung and Polani [2006] that provides kernel-based extensions of LSTD/LSPI, and Taylor and Parr [2009] that unifies some regularization-based extension of LSTD. Also see the proto-value function-based approach of Mahadevan and Maggioni [2007] and iLSTD of Geramifard et al. [2007].

2.4 Performance Loss Measures

Theoretical guarantees on the RL/Planning algorithm's performance are quantified by various performance loss measures. Two common families of performance loss measures are:

- Value function error $\|V^* - V^\pi\|_{p,\rho}$
- Online regret

Value Error

To understand the value function error as the performance loss, consider an RL/Planning algorithm that outputs a policy π. This policy might be the greedy policy w.r.t. an estimated action-value function $\hat{Q}$, that is $\pi = \hat{\pi}(\cdot; \hat{Q})$. Now suppose the agent starts at a specified initial state $X = x$ and follows the policy π. Its expected return would be $V^\pi(x)$. Comparing it with the expected return of following an optimal policy, $V^*(x)$, there will be a difference $V^*(x) - V^\pi(x) \geq 0$. If instead of starting from a fixed state $X = x$, the agent's initial state is distributed according to the "performance measuring" distribution $\rho \in \mathcal{M}(\mathcal{X})$, we may use the $L_1(\rho)$-norm of this error to determine the expected difference between the value of following policy π instead of π^*. The user-chosen probability distribution ρ reflects the

importance of various regions of the state space as the initial state distribution of the agent according to the user.

One may also extend this idea to other L_p-norms ($1 \leq p \leq \infty$) too. The L_1-norm has the interpretation we just described and is a natural choice. Another common choice is to use the L_∞-norm. This norm, however, is too pessimistic as a large point-wise performance loss error in a tiny subset of the state space leads to a large overall performance loss. This is not usually the type of result one would expect. Moreover, one may define the performance loss w.r.t. the action-value functions, i.e., $\|Q^* - Q^\pi\|_{p,\rho}$. This measures the performance loss whenever the initial action-state is selected according to $\rho \in \mathcal{M}(\mathcal{X} \times \mathcal{A})$.

Regret

Regret is another measure of performance loss for RL/Planning algorithms in online scenarios. It is usually defined as the expected difference between the return of the algorithm to the average optimal reward of the MDP (refer to Auer et al. [2009]; Bartlett and Tewari [2009]; Jaksch et al. [2010] and references therein). Currently, such regret guarantees are only available for *finite* state-action MDPs.

In this work, we focus on the value function error with the $L_p(\rho)$-norm as the measure of performance loss.

2.5 Reinforcement Learning and Planning in Large State Spaces

The use of function approximation in the value-based approaches to solve RL/Planning with large state space is inevitable in most cases. This topic has attracted the attention of many researchers in the past two decades. Without attempting to provide an extensive literature survey on different ways FA has been used in RL/Planning problems, we discuss two key aspects of various methods and provide some exemplar references. For extensive surveys of this topic, refer to Szepesvári [2010]; Bertsekas [2010]; Buşoniu et al. [2010a].

We categorize the use of FA in RL/Planning according to

- The modeling assumptions: Parametric vs. Nonparametric
- Statistical convergence guarantee

In the followings, we discuss these issues in detail.

The Modeling Assumption: Parametric vs. Nonparametric

Parametric Models

In the parametric approach to the value function estimation, the "structural" properties of the FA is set *a priori* and do not change according to data. Examples of these structural properties are the number of basis functions and their corresponding shape and place in a general linear model. We refer to this widely-used class of parametric model as *linear FA*, though one should be careful that the term *linear* denotes different concepts in other contexts. For instance, *linear estimation* is defined as when the estimate is described by a linear operator on target values; or linear is used in the approximation theory to refer to the concept of approximating a function within a linear span of a set of orthonormal basis for a given function space [Devore, 1998].

Linear FA

The use of linear FA to represent the value function is a common practice in the RL community. It has been applied both in the incremental [Sutton and Barto, 1998, Chapter 8] and the batch [Lagoudakis and Parr, 2003] algorithms, and their statistical properties are well-studied [Tsitsiklis and Van Roy, 1997; Sutton et al., 2009; Maillard et al., 2010]. Nevertheless, as argued in Section 1.1, whenever the true value function cannot be well-represented by the parametric model, no matter how elegant the value function estimation algorithm is, we have the function approximation error. This error might result in poor performance.

Finding the proper parametric model for a given problem is usually difficult and requires extensive trial and error. The proper choice of the parametric model depends on some

properties of the underlying problem and data, such as the number of available data samples, the geometry of data in the input space, and the smoothness of the target, which in general are unknown a priori.

Nonparametric Models

An alternative is the family of nonparametric approaches, which has been successful in the statistics and the supervised learning communities for a long time [Györfi et al., 2002; Hastie et al., 2001; Bishop, 2006; Wasserman, 2007]. These methods have weaker assumptions on the statistical model of the [value] function. They change the FA according to data, and upon the proper choice of their parameters, are adaptive to the problem in hand.

In the following, we first review some methods that are not truly nonparametric but share some similarities with them and then discuss methods that can be considered as nonparametric algorithms for RL/Planning problems.

Basis Adaptation

The basis adaptation approaches, which are not usually formulated in a truly nonparametric framework, work by parameterizing basis functions (e.g., the centers and the width of Radial Basis Functions) and fine-tuning these parameters to optimize an objective function such as an estimate of the Bellman residual error. For example, Menache et al. [2005] introduce a gradient-based method and the cross-entropy algorithm to find basis parameters that minimize an estimate of the Bellman residual error $\|V(\cdot;\theta) - T^\pi V(\cdot;\theta)\|_{\mathcal{X}'}$, in which $\mathcal{X}' \subset \mathcal{X}$ is a finite subset of $\mathcal{X}$ and θ in $V(\cdot;\theta)$ describes the parameters of basis functions. Yu and Bertsekas [2009] extend this idea to nonlinear T^*. These approaches are not nonparametric because they work with finite dimensional function spaces, but the use of nonlinear FA and data-dependent adaptation make them similar to many nonparametric methods.

Basis Generation

A nonparametric approach to solve RL/Planning problems is to generate new basis functions data and problem-dependently – as opposed to using a fixed pre-defined set of basis functions. The generated basis functions can then be used in any algorithm with linear FA. Basis generation can be done in different ways. One general approach is to benefit from some intrinsic properties of the MDP or the induced Markov chain, such as the transition probability kernel P and the reward function r, to build basis functions. For instance, one method is to use the set of eigenfunctions of P^π, that is $\{\rho_i : \rho_i P^\pi = \lambda_i \rho_i\}$, as basis functions. An extension of this method is to use the union of that set with $\{(P^\pi)^k r : k = 1, \ldots\}$, which leads to the so-called *augmented Krylov* method. These two methods have been suggested by Petrik [2007], who studies their approximation properties.

Another basis generation approach is to use the Bellman residual for defining new basis functions [Parr et al., 2007]. This approach starts from a single arbitrary basis function, and then estimates the value function $\hat{V}$. If the estimated value function is not the same as the true value function (because of both the estimation and the function approximation error), $\hat{V} - T^\pi \hat{V}$ will be a nonzero function called the Bellman residual. This residual defines a new basis function. It can be shown that if we ignore the estimation error, repeating this procedure decreases an upper bound on the function approximation error. Parr et al. [2008] show that if we start from r as the basis function for the Bellman residual basis function generation method [Parr et al., 2007], the result is the same as the Krylov basis $\{(P^\pi)^k r : k = 1, \ldots\}$ of Petrik [2007].

One must be careful in interpreting the aforementioned results. The theoretical guarantees on decreasing the upper bound on the approximation error are valid whenever we precisely find the eigenfunctions of P^π, functions $(P^\pi)^k r$, or the effect of the Bellman operator T^π on $\hat{V}$. Even if we know the model, these computations may be intractable for large MDPs. Moreover, if we use sample-based approaches to estimate these quantities, as suggested by Parr et al. [2007] for the estimation of the Bellman residual, it is not evident that the new auxiliary estimation problem is any easier than the original problem of estimating the value function itself.

Another similar basis generation method is a graph Laplacian-based approach [Mahadevan and Maggioni, 2007]. This method generates basis functions in accordance with the transition flow's geometry of the MDP. This choice might be helpful when the geometry of

most probable states has some special properties like lying close to a low-dimensional manifold. In this method, basis functions are eigenfunctions of the graph Laplacian operator. The graph Laplacian operator is built based on the state transition data and its spectrum contains information about the geometry of the transition flow in the state space [Chung, 1997]. This method, as opposed to the augmented Krylov method of Petrik [2007], does not take into account the reward function. Some may consider this as an advantage because of the transferability of basis functions over problems with the same dynamics but with different reward functions, whereas others may consider it as a disadvantage since not all available information has been used [Mahadevan and Maggioni, 2007].

Gaussian Process Temporal Difference (GPTD) is an example of nonparametric method to represent the value function [Engel et al., 2005]. In GPTD, one puts a GP prior over value function V^π. Define the residuals as $\Delta V^\pi(x) \triangleq V^\pi(x) - G_\gamma(\xi(x))$ with the trajectory $\xi(x)$ being the result of following policy π. By assuming that 1) $\Delta V^\pi(x)$ is a GP, and 2) $\Delta V^\pi(x_1)$ and $\Delta V^\pi(x_2)$ are independent for $x_1 \neq x_2$, one obtains a closed-form solution for the posterior of the value function given the observed data samples. GPTD, like many other RKHS-based machine learning algorithms, uses data to generate a dictionary of basis functions. GPTD is an example of nonparametric methods for the policy evaluation and GPSARSA is its modification to handle policy improvement. Nevertheless, because of the aforementioned assumptions on the probabilistic model underlying residuals, GPTD lacks a firm theoretical justification.

Nonparametric AVI

As some instances of nonparametric and data-dependent approaches in the context of AVI, we mention Ormoneit and Sen [2002] who use smoothing kernel-based estimator, Ernst et al. [2005] who devise tree-based methods to represent the value function, and Buşoniu et al. [2010b] who use Fuzzy rule set-based FA in the inner loop of AVI. We would like to mention Riedmiller [2005] who applies neural networks and Lange and Riedmiller [2010] who utilize deep neural networks in the context of AVI. These algorithms, however, are not nonparametric as the size of the neural networks is fixed a priori. If these methods allowed the topology of the neural network to change as new data arrives, they could be considered as nonparametric methods too.

Regularization

An important class of nonparametric approaches is those that use a *regularization functional* (also called *regularizer* or *penalizer*) to control the complexity of a large function space. Even though the regularization technique has been a successful approach in the supervised learning literature for many decades [Hoerl and Kennard, 1970; Wahba, 1990; Tibshirani, 1996; Vapnik, 1998], its application in RL/Planning has been quite recent. Some exemplar papers are Engel et al. [2005]; Jung and Polani [2006]; Loth et al. [2007]; Farahmand et al. [2008, 2009a,b]; Taylor and Parr [2009]; Kolter and Ng [2009]. With the exception of Farahmand et al. [2008, 2009a,b], the other aforementioned examples do not provide any statistical guarantee on the performance of their algorithms.

RFQI (an AVI algorithm that is introduced in Chapter 5), and REG-LSPI and REG-BRM (API algorithms that are introduced in Chapter 6), are instances of regularization-based nonparametric methods for RL/Planning problems. If we formulate them as an optimization problem in an RKHS, they automatically generate basis functions to represent the action-value function. In contrast to the basis adaptation/generation algorithms where the basis generation is separated from the value function estimation, the basis generation procedure is an integral part of our value function estimation methods. If we formulate them in a function space with an over-complete dictionary or a Besov space with wavelet basis, the l_1-regularization-based methods may select a sparse subset of basis functions that is required for the value function estimation.

Statistical Convergence Guarantee

The convergence behavior of an algorithm shows how the agent performs after a certain amount of interactions with the environment. The convergence property of an algorithm can be stated by proving its *consistency* or *convergence rate* or other similar notions. Some

algorithms may not eventually converge, but still get close to the neighborhood of the "solution". They may still perform well, but not optimally.

As some examples of the statistical convergence guarantee for the MDPs with finite number of states and actions, Jaakkola et al. [1994] prove the asymptotic convergence of the action-value function estimates of the Q-learning algorithm to the optimal action-value function Q^*; Szepesvári [1997a] provides the asymptotic rate of convergence for the Q-learning algorithm; and Even-Dar and Mansour [2003] prove a finite-sample convergence rate.

The analysis is considerably more challenging for large state spaces and when FA is used to represent the [action-]value function. In these scenarios, asymptotic results are more common. Tsitsiklis and Van Roy [1997] proved that for the policy evaluation problem the Temporal Difference (TD) algorithm with linear FA provides a sequence of estimated value functions that converges to a close, though not diminishing, neighborhood of the projection of the true value function onto the span of basis functions. This result has an important restrictive assumption that the distribution of samples induced by the behavior policy is the same as the distribution that would be induced by the policy being evaluated (*target* policy). This scenario is known as the *on-policy* sampling. As a result, the result of Tsitsiklis and Van Roy [1997] does not hold for an algorithm such as Q-learning in which the target policy is different from the behavior policy (*off-policy* sampling scenario). Moreover, the result of Tsitsiklis and Van Roy [1997] is asymptotic and does not show either the finite sample behavior or convergence rate of the algorithm. In addition, due to the parametric representation of FA, the solution of TD with linear FA does not necessary get close to the true value function – it is not consistent in the usual statistical sense.

Extension of this result to *control*, in which a policy improvement is performed, has been an open problem for years. This is especially more difficult in the off-policy sampling scenario. Melo et al. [2008] prove that, under rather restrictive assumption, SARSA algorithm, which is an incremental online on-policy value iteration algorithm, with linear FA converges to the fixed point of a modified Bellman optimality operator defined by de Farias and Van Roy [2000]. They show that the algorithm does not behave erratically, which is not uncommon in RL/Planning with FA.

More recently, Sutton et al. [2009] address the problem of policy evaluation with off-policy sampling and show the asymptotic convergence of a modified TD algorithm with linear FA. Maei et al. [2009] extend this work to nonlinear, but still parametric, FA. The essence of these work is to minimize an objective function, called the *Projected Bellman Error*, through a stochastic gradient descent-like procedure. Roughly speaking, because these algorithms are gradient-based, upon the appropriate choice of step sizes, their convergence is guaranteed. The Projected Bellman Error objective function is the same as the way the LSTD loss function is described by Antos et al. [2008b] and is similar to the one we suggest in REG-LSTD (Farahmand et al. [2009b] and Chapter 6). The difference with the latter is in our use of regularized objective function. If the regularization coefficient is set to zero, these two objective functions are the same. Maei et al. [2010] introduce a stochastic subgradient algorithm for the problem of control with off-policy data samples. They assumed that the FA is linear and the behavior policy is fixed. Under some technical assumptions, they show that the algorithm converges to a local minimum of the Projected Bellman operator.

Even though most results in the RL/Planning literature concern asymptotic convergence of algorithms, some papers study the algorithms' finite sample error upper bound and/or convergence rate in offline setting. Antos et al. [2008b] study the finite-sample error upper bounds of a modified Bellman Residual Minimization algorithm used in the API procedure. Their result is stated for continuous state and discrete action spaces with a general form of function approximation. Munos and Szepesvári [2008] study the finite-sample error upper bound of Fitted Q-Iteration, an AVI algorithm, for continuous state and discrete action spaces. Antos et al. [2008a] study the same problem with *continuous* action space. These papers deal with the general choice of function spaces, so they can be considered as analysis of a potentially nonparametric method. They, however, do not concern how the function

space should be chosen. Also their result shows a suboptimal error upper bound. Maillard et al. [2010], on the other hand, study the Bellman Residual Minimization algorithm with linear FA. They assume that they have access to the generative model of the environment.

In Chapters 5 and 6 of this work, we introduce nonparametric AVI/API algorithms and provide finite-sample error upper bounds for them. The setup and the type of results of these chapters might be considered most similar to Munos and Szepesvári [2008]; Antos et al. [2008b]. The difference is that we focus on providing algorithms that use specific regularities of the problem. Moreover, the bounds in this work are considerably tighter than the previous similar results.

Chapter 3

Error Propagation for Approximate Policy and Value Iteration

3.1 Introduction

The exact solution of reinforcement learning and planning problems with large state space is difficult or impossible to obtain, so one usually has to aim for approximate solutions (Section 2.5). Approximate Policy Iteration (API) and Approximate Value Iteration (AVI) are two classes of iterative algorithms to solve RL/Planning problems with large state spaces. They try to approximately find the fixed point of the Bellman optimality operator.[1]

AVI starts from an initial value function Q_0 (or V_0), and iteratively applies an *approximation* of the Bellman optimality operator T^* (or T^π for the policy evaluation problem) to the previous estimate, i.e., $Q_{k+1} \approx T^* Q_k$ at iteration k. In general, Q_{k+1} is not equal to $T^* Q_k$ because 1) we do not have direct access to the Bellman operator but only have some samples from it, and 2) the function space to which Q belongs might not be representative enough. Thus there would be an approximation error $\varepsilon_k = T^* Q_k - Q_{k+1}$ between the result of the exact VI and AVI.[2]

API is another iterative algorithm to find an approximate solution to the fixed point of the Bellman *optimality* operator. It starts from a policy π_0, and then approximately *evaluates* that policy π_0, i.e., it finds a Q_k that satisfies $T^{\pi_k} Q_k \approx Q_k$ at iteration k. Afterwards, it performs a policy improvement step, which is to calculate the greedy policy w.r.t. the most recent action-value function, to get a new policy π_1, i.e., $\pi_{k+1}(\cdot) = \arg\max_{a \in \mathcal{A}} Q_k(\cdot, a)$ at iteration k. The policy iteration algorithm continues by approximately evaluating the newly obtained policy π_1 to get Q_1 and repeating the whole process again, generating a sequence of policies and their corresponding approximate action-value functions: $Q_0 \to \pi_1 \to Q_1 \to \pi_2 \to \cdots$. Similar to AVI, we may encounter a difference between the approximate solution Q_k ($T^{\pi_k} Q_k \approx Q_k$) and the true action-value of the policy Q^{π_k}, which is the solution of the fixed-point equation $T^{\pi_k} Q^{\pi_k} = Q^{\pi_k}$. Two convenient ways to describe this error are the Bellman residual of Q_k ($\varepsilon_k = Q_k - T^{\pi_k} Q_k$) and the policy evaluation approximation error ($\varepsilon_k = Q_k - Q^{\pi_k}$).

A crucial question in the applicability of API/AVI, which is the main topic of this chapter, is to understand how either the approximation error or the Bellman residual at each iteration of API/AVI affects the quality of the resulting policy. Suppose we run API/AVI

[1] This chapter is the result of the collaboration of the author with Rémi Munos and Csaba Szepesvári.

[2] The notion of approximation error that we use in this chapter should not be confused with the term *function approximation error* used in the statistical learning theory. Here we are merely referring to the error caused by approximately performing VI/PI.

for K iterations to obtain a policy π_K. Does the knowledge that all $(\varepsilon_k)_{k=0}^{K-1}$ are small (maybe because we have had a lot of samples and used powerful function approximators) imply that V^{π_K} is close to the optimal value function V^* too? If so, how does the error occurred at a certain iteration k **propagate** through iterations of API/AVI and affect the final performance loss?

There have already been some results that partially address this question. As an example, Proposition 6.2 of Bertsekas and Tsitsiklis [1996] shows that for API applied to a finite MDP, we have

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

Similarly for AVI, if the approximation errors are uniformly bounded, that is $\|T^*V_k - V_{k+1}\|_\infty \leq \varepsilon$, we have [Munos, 2007]

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

Nevertheless, most of these results are pessimistic in several ways. To begin with, they are expressed as the supremum norm of the approximation errors $\|Q^{\pi_k} - Q_k\|_\infty$ or the Bellman error $\|Q_k - T^{\pi_k}Q_k\|_\infty$. The supremum norm is conservative compared to the L_p-norms. It is quite possible that the error ε_k of a learning algorithm has a small L_p-norm, but a large L_∞-norm. It is desirable to have a result expressed by the L_p-norm of the approximation/Bellman residual ε_k.

In the recent past, there have been some attempts to extend the L_∞-norm results to the L_p ones [Munos, 2003, 2007; Antos et al., 2008b]. As a typical example, we quote the following from Antos et al. [2008b]:

Proposition 3.1 (Error Propagation for API – Antos et al. [2008b]). *Let $p \geq 1$ be a real and K be a positive integer and $\nu, \rho \in \mathcal{M}(\mathcal{X})$. Then, for any sequence of functions $(Q^{(k)})_{k=0}^{K-1} \subset B(\mathcal{X}\times\mathcal{A}; Q_{max})$, and their corresponding Bellman residuals $\varepsilon_k = Q_k - T^\pi Q_k$, the following inequality holds:*

$$\|Q^* - Q^{\pi_K}\|_{p,\rho} \leq \frac{2\gamma}{(1-\gamma)^2}\Big(C_{\rho,\nu}^{1/p} \max_{0\leq k<K} \|\varepsilon_k\|_{p,\nu} + \gamma^{\frac{K}{p}-1} R_{max}\Big),$$

where R_{max} is an upper bound on the magnitude of the expected reward function and

$$C_{\rho,\nu} = (1-\gamma)^2 \sum_{m\geq 1} m\gamma^{m-1} \sup_{\pi_1,\ldots,\pi_m} \left\| \frac{d\left(\rho P^{\pi_1}\cdots P^{\pi_m}\right)}{d\nu} \right\|_\infty .$$

The choice of ρ and ν in this and all further results is arbitrary, however, a natural choice for ν is the sampling distribution of the data, which is used by the policy evaluation module. On the other hand, the probability distribution ρ reflects the importance of various regions of the state space and is selected by the user.

This result indeed uses the $L_p(\nu)$-norm of the Bellman residuals and is an improvement over results like Bertsekas and Tsitsiklis [1996, Proposition 6.2], but still is pessimistic in some other ways and does not answer several important questions. For instance, this result implies that the uniform-over-all-iterations upper bound $\max_{0\leq k<K} \|\varepsilon_k\|_{p,\nu}$ is the quantity that determines the performance loss. One may wonder if this condition is really necessary, and ask whether it is better to put more emphasis on earlier/later iterations? Or another question is whether the appearance of terms in the form of $\|\frac{d(\rho P^{\pi_1}\cdots P^{\pi_m})}{d\nu}\|_\infty$ is intrinsic to the difficulty of the problem or can be relaxed.

The goal of this work is to answer these questions and to provide tighter upper bounds on the performance loss of API/AVI algorithms. These bounds help one understand what factors contribute to the difficulty of a learning problem. We base our analysis on the work of Munos [2007]; Antos et al. [2008b]; Munos [2003] and provide upper bounds on the performance loss in the form of $\|Q^* - Q^{\pi_k}\|_{1,\rho}$ (the expected loss weighted according

to the evaluation probability distribution ρ) for API (Section 3.2) and AVI (Section 3.3). This performance loss depends on a certain function of the ν-weighted L_2-norms of $(\varepsilon_k)_{k=0}^{K-1}$, in which ν is the data sampling distribution, and $C_{\rho,\nu}(K)$ is a function of the MDP, two probability distributions ρ and ν, and the number of iterations K (the results are more general and apply to the L_p-norms).

In addition to relating the performance loss to the L_p-norm of the Bellman residual/approximation error, this work has three main contributions that to our knowledge have not been considered before: 1) We show that the performance loss depends on the *expectation* of the squared Radon-Nikodym derivative of a certain distribution, to be specified in Section 3.2, rather than its supremum, as suggested by Munos [2003, 2007]; Antos et al. [2008b]. The difference between this expectation and the supremum can be considerable. For instance, for a finite state space with N states, the ratio can be of order $O(N^{1/2})$. 2) The contribution of the Bellman/approximation error to the performance loss is more prominent in later iterations of API/AVI and the effect of an error term in early iterations decays exponentially fast. 3) There are certain structures in the definition of concentrability coefficients that have not been explored before. We thoroughly discuss these *qualitative/structural* improvements in Section 3.4.

3.2 Approximate Policy Iteration

Consider the API procedure and the sequence $Q_0 \to \pi_1 \to Q_1 \to \pi_2 \to \cdots \to Q_{K-1} \to \pi_K$, where π_k is the greedy policy w.r.t. Q_{k-1} and Q_k is the approximate action-value function for policy π_k. For the sequence $(Q_k)_{k=0}^{K-1}$, denote the **B**ellman **R**esidual (BR) and the policy **A**pproximation **E**rror (AE) at each iteration by

$$\varepsilon_k^{\mathrm{BR}} \triangleq Q_k - T^{\pi_k} Q_k, \tag{3.1}$$

$$\varepsilon_k^{\mathrm{AE}} \triangleq Q_k - Q^{\pi_k}. \tag{3.2}$$

The goal of this section is to study the effect of ν-weighted L_{2p}-norm of the Bellman residual sequence $(\varepsilon_k^{\mathrm{BR}})_{k=0}^{K-1}$ or the policy evaluation approximation error sequence $(\varepsilon_k^{\mathrm{AE}})_{k=0}^{K-1}$ on the performance loss $\|Q^* - Q^{\pi_K}\|_{p,\rho}$ of the resulting policy π_K. We see that some intrinsic properties of the MDP affect the resulting bound. The main result of this section is stated as Theorem 3.2.

Due to the dynamical nature of 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 of the form $\rho P^{\pi_1} P^{\pi_2} \cdots$. The precise form of this dependence will be formalized in Theorem 3.2 (for API) and Theorem 3.4 (for AVI). Before stating the results, we shall define the following *concentrability* coefficients, which are relaxed version of those defined by Munos [2003, 2007]; Antos et al. [2008b].

Definition 3.1 (Expected Concentrability of the Future State-Action Distribution)**.** *Given $\rho, \nu \in \mathcal{M}(\mathcal{X} \times \mathcal{A})$, $m \geq 0$, and an arbitrary sequence of stationary policies $(\pi_m)_{m \geq 1}$, let $\rho P^{\pi_1} P^{\pi_2} \cdots P^{\pi_m} \in \mathcal{M}(\mathcal{X} \times \mathcal{A})$ 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 $(\pi_k)_{k=1}^{m}$. For integers $m_1, m_2 \geq 1$ and policies π, π_1, π_2, define the following concentrability*

coefficients, which are used in the analysis of API:

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

$$c_{PI_2,\rho,\nu}(m_1,m_2;\pi_1,\pi_2) \triangleq \left(\mathbb{E}\left[\left|\frac{d\left(\rho(P^{\pi^*})^{m_1}(P^{\pi_1})^{m_2}P^{\pi_2}\right)}{d\nu}(X,A)\right|^2\right]\right)^{\frac{1}{2}},$$

$$c_{PI_3,\rho,\nu} \triangleq \left(\mathbb{E}\left[\left|\frac{d\left(\rho P^{\pi^*}\right)}{d\nu}(X,A)\right|^2\right]\right)^{\frac{1}{2}},$$

with $(X,A) \sim \nu$. *If the future state-action distribution* $\rho(P^{\pi^*})^{m_1}(P^{\pi})^{m_2}$ *(or* $\rho(P^{\pi^*})^{m_1}(P^{\pi_1})^{m_2}P^{\pi_2}$ *or* ρP^{π^*}*) is not absolutely continuous w.r.t.* ν*, then we take* $c_{PI_1,\rho,\nu}(m_1,m_2;\pi)=\infty$ *(and similarly for others). Also for integers* $m_1,m_2\geq 1$*, policy* π *and the sequence of policies* $\pi_1,\ldots,\pi_k$ *define the following concentrability coefficient, which are used in the analysis of AVI:*

$$c_{VI_1,\rho,\nu}(m_1,m_2;\pi) \triangleq \left(\mathbb{E}\left[\left|\frac{d\left(\rho(P^{\pi})^{m_1}(P^{\pi^*})^{m_2}\right)}{d\nu}(X,A)\right|^2\right]\right)^{\frac{1}{2}},$$

$$c_{VI_2,\rho,\nu}(m_1;\pi_1,\ldots,\pi_k) \triangleq \left(\mathbb{E}\left[\left|\frac{d\left(\rho(P^{\pi_k})^{m_1}P^{\pi_{k-1}}P^{\pi_{k-2}}\cdots P^{\pi_1}\right)}{d\nu}(X,A)\right|^2\right]\right)^{\frac{1}{2}},$$

with $(X,A)\sim\nu$. *If the future state-action distribution* $\rho(P^{\pi})^{m_1}(P^{\pi^*})^{m_2}$ *(or likewise* $\rho(P^{\pi_k})^{m_1}P^{\pi_{k-1}}P^{\pi_{k-2}}\cdots P^{\pi_1}$*) is not absolutely continuous w.r.t.* ν*, then we take* $c_{VI_1,\rho,\nu}(m_1,m_2;\pi)=\infty$ *(similarly,* $c_{VI_2,\rho,\nu}(m_1;\pi_1,\ldots,\pi_k)=\infty$*).*

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

$$\alpha_k = \begin{cases} \frac{(1-\gamma)\gamma^{K-k-1}}{1-\gamma^{K+1}} & 0\leq k<K, \\ \frac{(1-\gamma)\gamma^{K}}{1-\gamma^{K+1}} & k=K. \end{cases} \tag{3.3}$$

Theorem 3.2 (Error Propagation for API)**.** *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* $(Q_k)_{k=0}^{K-1}\subset B(\mathcal{X}\times\mathcal{A},Q_{max})$ *and the corresponding sequence* $(\varepsilon_k)_{k=0}^{K-1}$ *defined in* (3.1) *or* (3.2)*, we have*[3]

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

where $\mathcal{E}(\varepsilon_0,\ldots,\varepsilon_{K-1};r)=\sum_{k=0}^{K-1}\alpha_k^{2r}\|\varepsilon_k\|_{2p,\nu}^{2p}$ *and* $C_{PI(BR/AE),\rho,\nu}$ *is defined below.*
(a) If $\varepsilon_k=\varepsilon^{BR}$ *for all* $0\leq k<K$*, we have*

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

[3]The proof actually shows a bound that is tighter than the statement of the theorem, but we simplified it to be more accessible.

(b) If $\varepsilon_k = \varepsilon^{AE}$ for all $0 \le k < K$, we have

$$C_{PI(AE),\rho,\nu}(K;r) = \left(\frac{1-\gamma}{2}\right)^2 \sup_{\pi'_0,\ldots,\pi'_K} \sum_{k=0}^{K-1} \alpha_k^{2(1-r)} \left(\sum_{m\ge 0} \gamma^m c_{PI_1,\rho,\nu}(K-k-1, m+1; \pi'_{k+1}) + \sum_{m\ge 1} \gamma^m c_{PI_2,\rho,\nu}(K-k-1, m; \pi'_{k+1}, \pi'_k) + c_{PI_3,\rho,\nu} \right)^2.$$

Proof. Part (a): Let $E_k = P^{\pi_{k+1}}(\mathbf{I} - \gamma P^{\pi_{k+1}})^{-1} - P^{\pi^*}(\mathbf{I} - \gamma P^{\pi_k})^{-1}$. It can be shown that (Munos [2003, Lemma 4])

$$Q^* - Q^{\pi_{k+1}} \le \gamma P^{\pi^*}(Q^* - Q^{\pi_k}) + \gamma E_k \varepsilon_k^{\mathrm{BR}}.$$

By induction, we get

$$Q^* - Q^{\pi_K} \le \gamma \sum_{k=0}^{K-1} (\gamma P^{\pi^*})^{K-k-1} E_k \varepsilon_k^{\mathrm{BR}} + (\gamma P^{\pi^*})^K (Q^* - Q^{\pi_0}). \tag{3.4}$$

Define $F_k = P^{\pi_{k+1}}(\mathbf{I} - \gamma P^{\pi_{k+1}})^{-1} + P^{\pi^*}(\mathbf{I} - \gamma P^{\pi_k})^{-1}$, and take point-wise absolute value of (3.4) to get

$$|Q^* - Q^{\pi_K}| \le \gamma \sum_{k=0}^{K-1} (\gamma P^{\pi^*})^{K-k-1} F_k |\varepsilon_k^{\mathrm{BR}}| + (\gamma P^{\pi^*})^K |Q^* - Q^{\pi_0}|.$$

Since $\sum_{k=0}^K \alpha_k = 1$, for a convex function $\phi(\cdot)$ and a real-valued sequence $(f_k)_{k=0}^K$, Jensen's inequality $\phi(\sum_{k=0}^K a_k f_k) \le \sum_{k=0}^K a_k \phi(f_k)$ holds. Introduce the sequence $(A_k)_{k=0}^K$ to simplify our further analysis:

$$A_k = \begin{cases} \frac{1-\gamma}{2}(P^{\pi^*})^{K-k-1} F_k & 0 \le k < K, \\ (P^{\pi^*})^K & k = K. \end{cases}$$

It is shown in Lemma 12 of Antos et al. [2008b] that 1) $A_k : B(\mathcal{X}\times\mathcal{A}) \to B(\mathcal{X}\times\mathcal{A})$ are positive linear operators that satisfy $A_k \mathbf{1} = \mathbf{1}$, and 2) if $\phi(\cdot)$ is convex, then $\phi(A_k Q) \le A_k(\phi(Q))$ where ϕ is applied point-wise.

Using these notations and noting that $Q^* - Q^{\pi_0} \le \frac{2}{1-\gamma} R_{\max} \mathbf{1}$ (where $\mathbf{1}$ is the constant function defined on domain $\mathcal{X}\times\mathcal{A}$ with the value of 1), we get

$$|Q^* - Q^{\pi_K}| \le \frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2} \left[\sum_{k=0}^{K-1} \alpha_k A_k |\varepsilon_k^{\mathrm{BR}}| + \gamma^{-1} \alpha_K A_K R_{\max} \mathbf{1} \right]. \tag{3.5}$$

Denote $\lambda_K = [\frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2}]^p$. Take the p^{th} power of both sides of (3.5) and apply Jensen's inequality twice (once considering (A_k) and once considering (α_k)) to get

$$\begin{aligned} \|Q^* - Q^{\pi_K}\|_{p,\rho}^p &= \int_{\mathcal{X}\times\mathcal{A}} |Q^*(x,a) - Q^{\pi_K}(x,a)|^p \rho(dx) \\ &\le \lambda_K \rho \left[\sum_{k=0}^{K-1} \alpha_k A_k |\varepsilon_k^{\mathrm{BR}}|^p + \gamma^{-p} \alpha_K A_K R_{\max}^p \mathbf{1} \right]. \end{aligned}$$

Consider a term such as

$$\rho A_k |\varepsilon_k^{\mathrm{BR}}|^p = \frac{1-\gamma}{2} \rho (P^{\pi^*})^{K-k-1} \left[P^{\pi_{k+1}}(\mathbf{I} - \gamma P^{\pi_{k+1}})^{-1} + P^{\pi^*}(\mathbf{I} - \gamma P^{\pi_k})^{-1} \right] |\varepsilon_k^{\mathrm{BR}}|^p$$

for any $0 \leq k < K$. Expand $(\mathbf{I} - \gamma P^{\pi_{k+1}})^{-1}$ and $(\mathbf{I} - \gamma P^{\pi_k})^{-1}$ to have

$$\rho A_k |\varepsilon_k^{\mathrm{BR}}|^p = \frac{1-\gamma}{2}\rho \left[\sum_{m\geq 0} \gamma^m (P^{\pi^*})^{K-k-1} (P^{\pi_{k+1}})^{m+1} + \sum_{m\geq 0} \gamma^m (P^{\pi^*})^{K-k} (P^{\pi_k})^m \right] |\varepsilon_k^{\mathrm{BR}}|^p.$$

For any Borel measurable function $f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, and the probability measures μ_1 and μ_2 that satisfy $\mu_1 \ll \mu_2$, we have the following Cauchy-Schwarz inequality:

$$\int_{\mathcal{X}\times\mathcal{A}} f d\mu_1 \leq \left(\int_{\mathcal{X}\times\mathcal{A}} \left| \frac{d\mu_1}{d\mu_2} \right|^2 d\mu_2 \right)^{\frac{1}{2}} \left(\int_{\mathcal{X}\times\mathcal{A}} f^2 d\mu_2 \right)^{\frac{1}{2}}.$$

Let us focus on a single term such as $\rho(P^{\pi^*})^{K-k-1}(P^{\pi_{k+1}})^{m+1}|\varepsilon_k^{\mathrm{BR}}|^p$, and apply the Cauchy-Schwarz inequality to it. We have

$$\begin{aligned}
\rho(P^{\pi^*})^{K-k-1}(P^{\pi_{k+1}})^{m+1}|\varepsilon_k^{\mathrm{BR}}|^p &\leq \left(\int_{\mathcal{X}\times\mathcal{A}} \left| \frac{d\left(\rho(P^{\pi^*})^{K-k-1}(P^{\pi_{k+1}})^{m+1}\right)}{d\nu} \right|^2 d\nu \right)^{\frac{1}{2}} \\
&\quad \times \left(\int_{\mathcal{X}\times\mathcal{A}} |\varepsilon_k^{\mathrm{BR}}|^{2p} d\nu \right)^{\frac{1}{2}} \\
&= c_{\mathrm{PI}_1,\rho,\nu}(K-k-1, m+1; \pi_{k+1}) \left\| \varepsilon_k^{\mathrm{BR}} \right\|_{2p,\nu}^p.
\end{aligned}$$

Doing the same for the other terms $(P^{\pi^*})^{K-k}(P^{\pi_k})^m$, and noting that $\rho A_K \mathbf{1} = \rho \mathbf{1} = 1$ implies that

$$\begin{aligned}
&\|Q^* - Q^{\pi_K}\|_{p,\rho}^p \leq \\
&\lambda_K \Bigg[\frac{1-\gamma}{2} \sum_{k=0}^{K-1} \alpha_k \sum_{m\geq 0} \gamma^m \left(c_{\mathrm{PI}_1,\rho,\nu}(K-k-1, m+1; \pi_{k+1}) + c_{\mathrm{PI}_1,\rho,\nu}(K-k, m; \pi_k) \right) \left\| \varepsilon_k^{\mathrm{BR}} \right\|_{2p,\nu}^p \\
&\quad + \gamma^{-p} \alpha_K R_{\max}^p \Bigg].
\end{aligned}$$

In order to separate concentrability coefficients and $(\varepsilon_k^{\mathrm{BR}})_{k=0}^{K-1}$, we use Hölder's inequality

$$\sum_{k=0}^{K-1} a_k b_k \leq \left(\sum_{k=0}^{K-1} |a_k|^s \right)^{\frac{1}{s}} \left(\sum_{k=0}^{K-1} |b_k|^{s'} \right)^{\frac{1}{s'}}$$

with $s \in (1, \infty)$ and $\frac{1}{s} + \frac{1}{s'} = 1$. Let $a_k = \alpha_k^r \left\| \varepsilon_k^{\mathrm{BR}} \right\|_{2p,\nu}^p$ and

$$b_k = \alpha_k^{1-r} \sum_{m\geq 0} \gamma^m \left(c_{\mathrm{PI}_1,\rho,\nu}(K-k-1, m+1; \pi_{k+1}) + c_{\mathrm{PI}_1,\rho,\nu}(K-k, m; \pi_k) \right)$$

for some $r \in [0, 1]$. Therefore for all $(s, r) \in (1, \infty) \times [0, 1]$, we have

$$\begin{aligned}
&\|Q^* - Q^{\pi_K}\|_{p,\rho}^p \leq \\
&\lambda_K \frac{1-\gamma}{2} \Bigg[\sum_{k=0}^{K-1} \alpha_k^{s(1-r)} \left(\sum_{m\geq 0} \gamma^m \left(c_{\mathrm{PI}_1,\rho,\nu}(K-k-1, m+1; \pi_{k+1}) + c_{\mathrm{PI}_1,\rho,\nu}(K-k, m; \pi_k) \right) \right)^s \Bigg]^{\frac{1}{s}} \\
&\quad \times \left[\sum_{k=0}^{K-1} \alpha_k^{s'r} \left\| \varepsilon_k^{\mathrm{BR}} \right\|_{2p,\nu}^{ps'} \right]^{\frac{1}{s'}} + \lambda_K \gamma^{-p} \alpha_K R_{\max}^p. \qquad (3.6)
\end{aligned}$$

Since $(\pi_k)_{k=0}^K$ are not known, we take the supremum over all policies. Moreover as (3.6) holds for all $(s,r) \in (1,\infty) \times [0,1]$, we may take the infimum over (s,r) in the right hand side. Also note that $\frac{1-\gamma}{1-\gamma^{K+1}} < 1$ and $\lambda_K \leq [\frac{2\gamma}{(1-\gamma)^2}]^p$. After taking the p^{th} root, we have

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

where

$$C_{\mathrm{PI(BR)},\rho,\nu}(K;r,s) =$$
$$\left(\frac{1-\gamma}{2}\right)^s \sup_{\pi'_0,\ldots,\pi'_K} \sum_{k=0}^{K-1} \alpha_k^{s(1-r)} \left(\sum_{m\geq 0} \gamma^m \left(c_{\mathrm{PI}_1,\rho,\nu}(K-k-1, m+1; \pi'_{k+1}) + c_{\mathrm{PI}_1,\rho,\nu}(K-k, m; \pi'_k) \right) \right)^s,$$

and $\mathcal{E}(\varepsilon_0^{\mathrm{BR}}, \ldots, \varepsilon_{K-1}^{\mathrm{BR}}; r, s) = \sum_{k=0}^{K-1} \alpha_k^{s'r} \left\|\varepsilon_k^{\mathrm{BR}}\right\|_{2p,\nu}^{ps'}$.

This result is general and holds for all $s \in (1,\infty)$. In order to make it more accessible, but at the cost of loosening of the upper bound, we simplify it by setting $s = s' = 2$. This finishes the proof of Part (a).

Part (b): The proof of this part is similar to the proof of Part (a). We briefly sketch the key steps: Define $E_k = P^{\pi_{k+1}}(\mathbf{I} - \gamma P^{\pi_{k+1}})^{-1}(\mathbf{I} - \gamma P^{\pi_k}) - P^{\pi^*}$. From Munos [2003, Lemma 4] one can show that

$$Q^* - Q^{\pi_K} \leq \gamma \sum_{k=0}^{K-1} (\gamma P^{\pi^*})^{K-k-1} E_k \varepsilon_k^{\mathrm{AE}} + (\gamma P^{\pi^*})^K (Q^* - Q^{\pi_0}). \tag{3.7}$$

Define $F_k = P^{\pi_{k+1}}(\mathbf{I} - \gamma P^{\pi_{k+1}})^{-1}(\mathbf{I} - \gamma P^{\pi_k}) + P^{\pi^*}$ and take the point-wise absolute value of (3.7) and use the same definition of A_k as Part (a) (with the new F_k) to get

$$|Q^* - Q^{\pi_K}| \leq \frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2} \left[\sum_{k=0}^{K-1} \alpha_k A_k |\varepsilon_k^{\mathrm{AE}}| + \gamma^{-1} \alpha_K A_K R_{\max} \mathbf{1} \right].$$

Consider a term like $\rho A_k |\varepsilon_k^{\mathrm{BR}}|^p$ for any $0 \leq k < K$ and expand $(\mathbf{I} - \gamma P^{\pi_{k+1}})^{-1}$. We have

$$\rho A_k |\varepsilon_k^{\mathrm{AE}}|^p = \frac{1-\gamma}{2} \rho \left[\sum_{m\geq 0} \gamma^m (P^{\pi^*})^{K-k-1} (P^{\pi_{k+1}})^{m+1} (\mathbf{I} - \gamma P^{\pi_k}) + P^{\pi^*} \right] |\varepsilon_k^{\mathrm{AE}}|^p.$$

After performing the same change of measure argument and applying the Cauchy-Schwarz inequality, we get

$$\|Q^* - Q^{\pi_K}\|_{p,\rho}^p \leq$$
$$\lambda_K \left[\frac{1-\gamma}{2} \sum_{k=0}^{K-1} \alpha_k \left(\sum_{m\geq 0} \gamma^m c_{\mathrm{PI}_1,\rho,\nu}(K-k-1, m+1; \pi_{k+1}) + \sum_{m\geq 1} \gamma^m c_{\mathrm{PI}_2,\rho,\nu}(K-k-1, m; \pi_{k+1}, \pi_k) + c_{\mathrm{PI}_3,\rho,\nu} \right) \left\|\varepsilon_k^{\mathrm{AE}}\right\|_{2p,\nu}^p + \gamma^{-p} \alpha_K R_{\max}^p \right].$$

Application of Hölder's inequality with a similarly defined decomposition and then taking the supremum over policies leads to

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

where

$$C_{\mathrm{PI(AE)},\rho,\nu}(K;r,s) = \left(\frac{1-\gamma}{2}\right)^s \sup_{\pi'_0,\ldots,\pi'_K} \sum_{k=0}^{K-1} \alpha_k^{s(1-r)} \left[\sum_{m\geq 0} \gamma^m c_{\mathrm{PI}_1,\rho,\nu}(K-k-1, m+1; \pi'_{k+1}) + \sum_{m\geq 1} \gamma^m c_{\mathrm{PI}_2,\rho,\nu}(K-k-1, m; \pi'_{k+1}, \pi'_k) + c_{\mathrm{PI}_3,\rho,\nu} \right]^s$$

and $\mathcal{E}(\varepsilon_0^{\mathrm{AE}}, \ldots, \varepsilon_{K-1}^{\mathrm{AE}}; r, s) = \sum_{k=0}^{K-1} \alpha_k^{s'r} \left\|\varepsilon_k^{\mathrm{AE}}\right\|_{2p,\nu}^{ps'}$. □

We discuss this result in detail in Section 3.4. To provide an upper bound for $\|V^* - V^{\pi_K}\|_{p,\nu}$, we may use the following lemma.

Lemma 3.3. *For the probability measure $\rho \in \mathcal{M}(\mathcal{X}\times\mathcal{A})$ that is absolutely continuous w.r.t. the Lebesgue measure $\lambda \in \mathcal{M}(\mathcal{X}\times\mathcal{A})$, let $\rho_{\mathcal{X}} \in \mathcal{M}(\mathcal{X})$ denote its marginal on $\mathcal{X}$ and $\pi_\rho : \mathcal{X} \to \mathcal{M}(\mathcal{A})$ be its conditional probability conditioned on $x \in \mathcal{X}$. Define*

$$C_{2,\rho}^{Q\to V} \triangleq \left[\int_{\mathcal{X}} \max_{a\in\mathcal{A}} \left\{ \left| \frac{1}{\pi_\rho(a|x)} \right|^2 \right\} d\rho_{\mathcal{X}}(x) \right]^{1/2},$$
$$C_{\infty,\rho}^{Q\to V} \triangleq \sup_{x\in\mathcal{X}} \max_{a\in\mathcal{A}} \left\{ \frac{1}{\pi_\rho(a|x)} \right\}.$$

For an action-value function $Q \in \mathcal{F}^{|\mathcal{A}|}$, let π be the greedy policy w.r.t. Q, i.e, $\pi = \hat{\pi}(\cdot; Q)$. Then, we have

$$\|V^* - V^\pi\|_{1,\rho_{\mathcal{X}}} \leq \begin{cases} C_{2,\rho}^{Q\to V} \|Q^* - Q^\pi\|_{2,\rho}, \\ C_{\infty,\rho}^{Q\to V} \|Q^* - Q^\pi\|_{1,\rho}. \end{cases}$$

Proof. First note that $|V^*(x) - V^\pi(x)| = |\max_{a\in\mathcal{A}} Q^*(x,a) - \max_{a\in\mathcal{A}} Q^\pi(x,a)| \leq \max_{a\in\mathcal{A}} |Q^*(x,a) - Q^\pi(x,a)|$. Therefore, we have

$$\begin{aligned} &\int_{\mathcal{X}} d\rho_{\mathcal{X}}(x) |V^*(x) - V^\pi(x)| \leq \int_{\mathcal{X}} d\rho_{\mathcal{X}}(x) \max_{a\in\mathcal{A}} |Q^*(x,a) - Q^\pi(x,a)| \\ &\leq \int_{\mathcal{X}} d\rho_{\mathcal{X}}(x) \max_{a'\in\mathcal{A}} \left\{ \frac{1}{\pi_\rho(a'|x)} \right\} \sum_{a\in\mathcal{A}} \pi_\rho(a|x) |Q^*(x,a) - Q^\pi(x,a)| \\ &= \int_{\mathcal{X}\times\mathcal{A}} d\rho(x,a) \max_{a'\in\mathcal{A}} \left\{ \frac{1}{\pi_\rho(a'|x)} \right\} |Q^*(x,a) - Q^\pi(x,a)| \\ &\leq \begin{cases} \left[\int_{\mathcal{X}\times\mathcal{A}} d\rho(x,a) \left| \max_{a'\in\mathcal{A}} \left\{ \frac{1}{\pi_\rho(a'|x)} \right\} \right|^2 \right]^{1/2} \|Q^* - Q^\pi\|_{2,\rho}, \\ \sup_{x\in\mathcal{X}} \max_{a'\in\mathcal{A}} \left[\frac{1}{\pi_\rho(a'|x)} \right] \|Q^* - Q^\pi\|_{1,\rho}. \end{cases} \end{aligned}$$

where we used $\max_{a\in\mathcal{A}} |f(a)| \leq \max_{a'\in\mathcal{A}} \{\frac{1}{\pi_\rho(a|x)}\} \sum_{a\in\mathcal{A}} \pi_\rho(a|x) |f(a)|$ in the second inequality and applied the Cauchy-Schwarz inequality in the first case of the last inequality. We used the absolute continuity of ρ w.r.t. the Lebesgue measure λ to ensure that

the Radon-Nikodym derivate $\frac{d\rho}{d\lambda}$ existed, so we could decompose it as $(\frac{d\rho}{d\lambda})^{1/2}(\frac{d\rho}{d\lambda})^{1/2}$ and then apply the Cauchy-Schwarz inequality. To get $C_{2,\rho}^{Q\to V}$, note that for a positive f, $|\max_{a'\in\mathcal{A}} f(a)|^2 = \max_{a'\in\mathcal{A}} |f(a)|^2$, and marginalize the integral over $a \in \mathcal{A}$. □

This lemma indicates that in order to upper bound the value-based performance loss by an action-value-based performance loss, the probability distribution ρ should give enough weight to all actions over all states. Depending on whether the upper bound is expressed in the L_1 or L_2-norm of $Q^* - Q^\pi$, the multiplicative coefficient requires the supremum or the average value of $\max_a \frac{1}{\pi_\rho(a|x)}$ over x not to be large. Lemma 3.3 is independent of how Q is estimated and can be applied for both API and AVI.

Remark 3.1. One may be tempted to analyze the effect of the projected Bellman error $\varepsilon_k^{\text{PBR}} \triangleq Q_k - \Pi_{\nu,\mathcal{F}^{|\mathcal{A}|}} T^{\pi_k} Q_k$, with $\Pi_{\nu,\mathcal{F}^{|\mathcal{A}|}}$ being the ν-weighted projection operator onto the function space $\mathcal{F}^{|\mathcal{A}|}$ (Section 2.1), on the performance loss $\|Q^* - Q^{\pi_K}\|_{p,\rho}$. Unfortunately, the size of ε^{PBR} alone does not convey all necessary information about the closeness of Q to Q^π for it is possible that $\varepsilon^{\text{PBR}} = 0$ but $\|Q - Q^\pi\|_{p,\rho} > 0$.

One way to have a bound based on the projected Bellman error is to derive its corresponding Bellman error using the Pythagorean theorem alongside the function approximation error, i.e.,

$$\|Q - T^\pi Q\|^2 = \left\|\varepsilon^{\text{PBR}}\right\|^2 + \inf_{Q' \in \mathcal{F}^{|\mathcal{A}|}} \|Q' - T^\pi Q\|^2 ,$$

and then use results already proven for the Bellman error in this section.

3.3 Approximate Value Iteration

Consider the AVI procedure and the sequence of action-value function estimates $Q_0, Q_1, \ldots, Q_K$, in which Q_{k+1} is the result of approximately applying the Bellman optimality operator to the previous estimate Q_k, i.e., $Q_{k+1} \approx T^* Q_k$. Denote the approximation error caused at each iteration by

$$\varepsilon_k \triangleq T^* Q_k - Q_{k+1}. \tag{3.8}$$

The goal of this section is to analyze the AVI procedure and to relate the performance loss $\|Q^* - Q^{\pi_K}\|_{p,\rho}$ of the obtained policy $\pi_K(\cdot) = \hat{\pi}(\cdot; Q_K)$ (i.e., the greedy policy w.r.t. Q_K) to the approximation error sequence $(\varepsilon_k)_{k=0}^{K-1}$ and the properties of the MDP. The following theorem is the main result of this section.

Theorem 3.4 (Error Propagation for AVI)**.** *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 $(Q_k)_{k=0}^K \subset B(\mathcal{X} \times \mathcal{A}, Q_{max})$, and the corresponding sequence $(\varepsilon_k)_{k=0}^{K-1}$ defined in (3.8), we have*[4]

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

where

$$C_{VI,\rho,\nu}(K; r) = \left(\frac{1-\gamma}{2}\right)^2 \sup_{\pi'_1,\ldots,\pi'_K} \sum_{k=0}^{K-1} a_k^{2(1-r)} \left[\sum_{m\geq 0} \gamma^m \Big(c_{VI_1,\rho,\nu}(m, K-k; \pi'_K) + c_{VI_2,\rho,\nu}(m+1; \pi'_{k+1}, \ldots, \pi'_K) \Big) \right]^2 ,$$

and $\mathcal{E}(\varepsilon_0, \ldots, \varepsilon_{K-1}; r) = \sum_{k=0}^{K-1} \alpha_k^{2r} \|\varepsilon_k\|_{2p,\nu}^{2p}$.

[4]The proof actually shows a bound that is tighter than the statement of the theorem, but we simplified it to be more accessible.

Proof. First we derive a point-wise bound relating $Q^* - Q^{\pi_K}$ to $(\varepsilon_k)_{k=0}^{K-1}$ similar to Lemma 4.1 of Munos [2007]:

$$Q^* - Q_{k+1} = T^{\pi^*}Q^* - T^{\pi^*}Q_k + T^{\pi^*}Q_k - T^*Q_k + \varepsilon_k \leq \gamma P^{\pi^*}(Q^* - Q_k) + \varepsilon_k,$$
$$Q^* - Q_{k+1} = T^*Q^* - T^{\pi_k}Q^* + T^{\pi_k}Q^* - T^*Q_k + \varepsilon_k \geq \gamma P^{\pi_k}(Q^* - Q_k) + \varepsilon_k,$$

where we used the property of the Bellman optimality operator $T^*Q_k \geq T^{\pi^*}Q_k$, the definition of greedy policy π_k that entails $T^{\pi_k}Q_k = T^*Q_k$, and the definition of ε_k (3.8). By induction,

$$Q^* - Q_K \leq \sum_{k=0}^{K-1} \gamma^{K-k-1}(P^{\pi^*})^{K-k-1}\varepsilon_k + \gamma^K (P^{\pi^*})^K (Q^* - Q_0),$$
$$Q^* - Q_K \geq \sum_{k=0}^{K-1} \gamma^{K-k-1}(P^{\pi_{K-1}}P^{\pi_{K-2}}\cdots P^{\pi_{k+1}})\varepsilon_k + \gamma^K (P^{\pi_K}\cdots P^{\pi_1})(Q^* - Q_0). \quad (3.9)$$

Benefiting from $T^*Q_K \geq T^{\pi^*}Q_K$ and noting that $T^*Q_K = T^{\pi_K}Q^{\pi_K}$ by the definition of the greedy policy,

$$\begin{aligned} Q^* - Q^{\pi_K} &= T^{\pi^*}Q^* - T^{\pi^*}Q_K + T^{\pi^*}Q_K - T^*Q_K + T^*Q_K - T^{\pi_K}Q^{\pi_K} \\ &\leq T^{\pi^*}Q^* - T^{\pi^*}Q_K + T^*Q_K - T^{\pi_K}Q^{\pi_K} \\ &= \gamma P^{\pi^*}(Q^* - Q_K) + \gamma P^{\pi_K}(Q_K - Q^{\pi_K}) \\ &= \gamma P^{\pi^*}(Q^* - Q_K) + \gamma P^{\pi_K}(Q_K - Q^* + Q^* - Q^{\pi_K}). \end{aligned}$$

Re-arranging and using Lemma 4.2 of Munos [2007], we deduce that

$$Q^* - Q^{\pi_K} \leq \gamma(\mathbf{I} - \gamma P^{\pi_K})^{-1}(P^{\pi^*} - P^{\pi_K})(Q^* - Q_K). \quad (3.10)$$

Plugging (3.9) into (3.10) and taking the absolute value of both sides, we get the following point-wise inequality:

$$\begin{aligned} Q^* - Q^{\pi_K} \leq \gamma(\mathbf{I} - \gamma P^{\pi_K})^{-1} \Bigg[&\sum_{k=0}^{K-1} \gamma^{K-k-1} \left((P^{\pi^*})^{K-k} + (P^{\pi_K}P^{\pi_{K-1}}\cdots P^{\pi_{k+1}}) \right) |\varepsilon_k| \\ &+ \gamma^K \left((P^{\pi^*})^{K+1} + (P^{\pi_K}P^{\pi_K}P^{\pi_{K-1}}\cdots P^{\pi_1}) \right) |Q^* - Q_0| \Bigg]. \end{aligned} \quad (3.11)$$

As in the proof of Theorem 3.2, we use the sequence $(\alpha_k)_{k=0}^K$ defined in (3.3) and introduce

$$A_k = \begin{cases} \frac{1-\gamma}{2}(\mathbf{I} - \gamma P^{\pi_K})^{-1} \left[(P^{\pi^*})^{K-k} + (P^{\pi_K}P^{\pi_{K-1}}\cdots P^{\pi_{k+1}}) \right] & 0 \leq k < K, \\ \frac{1-\gamma}{2}(\mathbf{I} - \gamma P^{\pi_K})^{-1} \left((P^{\pi^*})^{K+1} + (P^{\pi_K}P^{\pi_K}P^{\pi_{K-1}}\cdots P^{\pi_1}) \right) & k = K. \end{cases}$$

Note that we use the same (α_k) as in the proof of Theorem 3.2, but (A_k) are different. Nevertheless, they satisfy the same properties that allow us to apply Jensen's inequality. By $|Q^* - Q_0| \leq \frac{2}{1-\gamma} R_{\max}\mathbf{1}$, we get

$$Q^* - Q^{\pi_K} \leq \frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2} \left[\sum_{k=0}^{K-1} \alpha_k A_k |\varepsilon_k| + \alpha_K A_K \frac{2}{1-\gamma} R_{\max}\mathbf{1} \right].$$

Now take the p^{th} power of both sides of (3.11), and apply Jensen inequality twice (once considering A_k and once considering α_k), to derive

$$\|Q^* - Q^{\pi_K}\|_{p,\rho}^p \leq \lambda_K \rho \left[\sum_{k=0}^{K-1} \alpha_k A_k |\varepsilon_k|^p + \alpha_K \left(\frac{2}{1-\gamma} \right)^p A_K R_{\max}^p \mathbf{1} \right].$$

Consider a term like $\rho A_k |\varepsilon_k|^p$ for any $0 \leq k < K$:

$$\begin{aligned}\rho A_k |\varepsilon_k|^p &= \frac{1-\gamma}{2}\rho(\mathbf{I}-\gamma P^{\pi_K})^{-1}\left[(P^{\pi^*})^{K-k} + (P^{\pi_K}P^{\pi_{K-1}}\cdots P^{\pi_{k+1}})\right]|\varepsilon_k|^p \\ &= \frac{1-\gamma}{2}\rho\left[\sum_{m\geq 0}\gamma^m\left((P^{\pi_K})^m(P^{\pi^*})^{K-k} + (P^{\pi_K})^{m+1}P^{\pi_{K-1}}\cdots P^{\pi_{k+1}})\right)\right]|\varepsilon_k|^p.\end{aligned}$$

Apply the Cauchy-Schwarz inequality, as we did in Theorem 3.2, to deduce

$$\begin{aligned}&\|Q^* - Q^{\pi_K}\|_{p,\rho}^p \leq \\ &\lambda_K\left[\frac{1-\gamma}{2}\sum_{k=0}^{K-1}\alpha_k\sum_{m\geq 0}\gamma^m\left(c_{\mathrm{VI}_1,\rho,\nu}(m,K-k;\pi_K) + c_{\mathrm{VI}_2,\rho,\nu}(m+1;\pi_{k+1},\ldots,\pi_K)\right)\|\varepsilon_k\|_{2p,\nu}^p\right. \\ &\left.+\alpha_K\left(\frac{2}{1-\gamma}\right)^p R_{\max}^p\right].\end{aligned}$$

Use Hölder's inequality with $a_k = \alpha_k^r \|\varepsilon_k\|_{2p,\nu}^p$ and

$$b_k = \alpha_k^{1-r}\sum_{m\geq 0}\gamma^m\left(c_{\mathrm{VI}_1,\rho,\nu}(m,K-k;\pi_K) + c_{\mathrm{VI}_2,\rho,\nu}(m+1;\pi_{k+1},\ldots,\pi_K)\right)$$

(all variables are defined the same as in the proof of Theorem 3.2). Therefore for all $(s,r) \in (1,\infty)\times[0,1]$, we have

$$\begin{aligned}&\|Q^* - Q^{\pi_K}\|_{p,\rho} \leq \\ &\frac{2\gamma}{(1-\gamma)^2}\left[\inf_{(s,r)\in(1,\infty)\times[0,1]} C_{\mathrm{VI},\rho,\nu}^{\frac{1}{ps}}(K;r,s)\mathcal{E}^{\frac{1}{ps'}}(\varepsilon_0,\ldots,\varepsilon_{K-1};r,s) + \frac{2}{1-\gamma}\gamma^{\frac{K}{p}}R_{\max}\right],\end{aligned}$$

where

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

and $\mathcal{E}(\varepsilon_0,\ldots,\varepsilon_{K-1};r,s) = \sum_{k=0}^{K-1}\alpha_k^{s'r}\|\varepsilon_k\|_{2p,\nu}^{ps'}$. To simplify the bound, at the cost of loosening the upper bound, we set $s = s' = 2$. □

Remark 3.2. One can obtain a similar upper bound on $\|V^* - V^{\pi_K}\|_{p,\rho_\mathcal{X}}$ when $(\varepsilon_k')_{k=0}^{K-1}$ are defined as $\varepsilon_k' \triangleq T^*V_k - V_{k+1}$ and $\rho_\mathcal{X}, \nu_\mathcal{X} \in \mathcal{M}(\mathcal{X})$. Denote $\rho_\mathcal{X}P^{\pi_1}P^{\pi_2}\cdots P^{\pi_m} \in \mathcal{M}(\mathcal{X})$ as the future-state distribution obtained when the first state is distributed according to $\rho_\mathcal{X}$ and then we follow the sequence of policies $(\pi_k)_{k=1}^m$. Define the following concentrability coefficients similar to Definition 3.1:

$$\begin{aligned}c_{\mathrm{VI}_1,\rho,\nu}'(m_1,m_2;\pi) &\triangleq \left(\mathbb{E}\left[\left|\frac{d\left(\rho_\mathcal{X}(P^\pi)^{m_1}(P^{\pi^*})^{m_2}\right)}{d\nu_\mathcal{X}}(X)\right|^2\right]\right)^{\frac{1}{2}}, \\ c_{\mathrm{VI}_2,\rho,\nu}'(m_1;\pi_1,\ldots,\pi_k) &\triangleq \left(\mathbb{E}\left[\left|\frac{d\left(\rho_\mathcal{X}(P^{\pi_k})^{m_1}P^{\pi_{k-1}}P^{\pi_{k-2}}\cdots P^{\pi_1}\right)}{d\nu_\mathcal{X}}(X)\right|^2\right]\right)^{\frac{1}{2}},\end{aligned}$$

with $X \sim \nu_\mathcal{X}$. Then the exact same result as Theorem 3.4 holds by replacing $c_{\mathrm{VI}_1,\rho,\nu}$ and $c_{\mathrm{VI}_2,\rho,\nu}$ with $c_{\mathrm{VI}_1,\rho,\nu}'$ and $c_{\mathrm{VI}_2,\rho,\nu}'$and using $\|\varepsilon_k'\|_{2p,\nu_\mathcal{X}}$ instead of $\|\varepsilon_k\|_{2p,\nu}$ in the definition of $\mathcal{E}(\varepsilon_0,\ldots,\varepsilon_{K-1};r,s)$ [Farahmand et al., 2010].

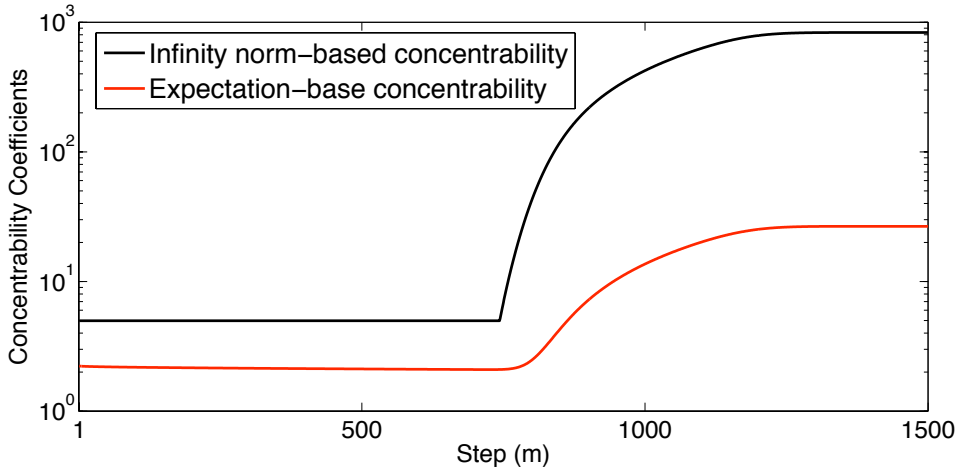

Figure 3.1: Comparison of $\left(\mathbb{E}\left[\left|\frac{d(\rho(P^\pi)^m)}{d\nu}\right|^2\right]\right)^{1/2}$ $(X \sim \nu)$ and $\left\|\frac{d(\rho(P^\pi)^m)}{d\nu}\right\|_\infty$. [The Y-scale is logarithmic.]

3.4 Discussion

In this section, we discuss significant improvements of Theorems 3.2 and 3.4 over previous results such as Bertsekas and Tsitsiklis [1996]; Munos [2003, 2007]; Antos et al. [2008b].

3.4.1 L_p-norm instead of L_∞-norm

As opposed to most error upper bounds, Theorems 3.2 and 3.4 relate $\|Q^* - Q^{\pi_K}\|_{p,\rho}$ (or $\|V^* - V^{\pi_K}\|_{p,\rho}$) to the L_p-norm of the approximation or Bellman errors $\|\varepsilon_k\|_{2p,\nu}$ of iterations in API/AVI. This should be contrasted with the traditional, and more conservative, results such as

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

for API (Proposition 6.2 of Bertsekas and Tsitsiklis [1996]). Refer to Section 2.4 for the discussion on why the L_p-norms are less conservative than the L_∞-norm.

Another benefit of the L_p-norm results compared to the L_∞ one is that the requirement of having a small L_p-norm error is usually less strict than the L_∞-norm's. The guarantee for the latter often requires strict conditions on the sampling distribution such as having a density bounded away from zero. This should be contrasted with the L_p-norm results that may even be distribution-free, i.e., they are not sensitive to the specific choice of the sampling distribution. The introduction of the L_p-norms for RL/Planning problems, however, is not new and has been done in the past couple of years in the work of Munos [2003, 2007]; Antos et al. [2008b] – see Proposition 3.1 in Section 3.1.

3.4.2 Expected versus supremum concentrability of the future state-action distribution

The concentrability coefficients (Definition 3.1) reflect the effect of future state-action distribution on the performance loss $\|Q^* - Q^{\pi_K}\|_{p,\rho}$. Previously it was thought that the key contributing factor to the performance loss is the supremum of the Radon-Nikodym derivative of these two distributions [Munos, 2003, 2007; Antos et al., 2008b]. This is evident in the definition of $C_{\rho,\nu}$ in Proposition 3.1 where we have terms in the form of $\|\frac{d(\rho(P^\pi)^m)}{d\nu}\|_\infty$ instead of $\left(\mathbb{E}\left[|\frac{d(\rho(P^\pi)^m)}{d\nu}(X,A)|^2\right]\right)^{1/2}$ (with $(X,A) \sim \nu$), which we have in Definition 3.1.

Nevertheless, it turns out that the key contributing factor that determines the performance loss is the *expectation* of the squared Radon-Nikodym derivative rather than its supremum. Intuitively this implies that even if for some subset of $\mathcal{X}' \times \mathcal{A}' \subset \mathcal{X} \times \mathcal{A}$ the ratio $\frac{d(\rho(P^\pi)^m)}{d\nu}$ is large but the probability $\nu(\mathcal{X}' \times \mathcal{A}')$ is very small, performance loss due to it is still small. This phenomenon has not been suggested by previous results.

As an illustration of this difference, consider a Chain Walk with 1000 states with a single policy that drifts toward state 1 of the chain. We start with $\rho(x) = \rho_{\mathcal{X}}(x) = \frac{1}{201}$ for $x \in [400, 600]$ and zero everywhere else. We then evaluate both $\|\frac{d(\rho(P^\pi)^m)}{d\nu}\|_\infty$ and $(\mathbb{E}\left[|\frac{d(\rho(P^\pi)^m)}{d\nu}|^2\right])^{1/2}$ $(X \sim \nu)$ for $m = 1, 2, \ldots$ when $\nu = \nu_{\mathcal{X}}$ is the uniform distribution over states. The result is shown in Figure 3.1. One observes that the ratio is constant in the beginning, but increases when the distribution $\rho(P^\pi)^m$ concentrates around state 1, until it reaches steady-state. The growth and the final value of the expectation-based concentrability coefficient is much smaller than that of the supremum-based coefficient.

It is easy to show that if the Chain Walk has N states and the policy has the same concentrating behavior and ν is uniform, then $\|\frac{d(\rho(P^\pi)^m)}{d\nu}\|_\infty \to N$, while for $X \sim \nu$, $(\mathbb{E}\left[|\frac{d(\rho(P^\pi)^m)}{d\nu}|^2\right])^{1/2} \to \sqrt{N}$ when $m \to \infty$. The ratio, therefore, would be of order $\Theta(\sqrt{N})$. This clearly shows the improvement of this new analysis in a simple problem. One may anticipate that this behavior occurs in many other problems too.

More generally, consider $C_\infty = \|\frac{d\mu}{d\nu}\|_\infty$ and $C_{L_2} = (\mathbb{E}\left[|\frac{d\mu}{d\nu}|^2\right])^{1/2}$ $(X \sim \nu)$. For a finite state space with N states and ν is the uniform distribution, $C_\infty \leq N$ but $C_{L_2} \leq \sqrt{N}$. Neglecting all other differences between our results and the previous ones, we get a performance upper bound in the form of $\|Q^* - Q^{\pi_K}\|_{1,\rho} \leq c_1(\gamma) O(N^{1/4}) \sup_k \|\varepsilon_k\|_{2,\nu}$, while Proposition 3.1 implies that $\|Q^* - Q^{\pi_K}\|_{1,\rho} \leq c_2(\gamma) O(N^{1/2}) \sup_k ||\epsilon_k||_{2,\nu}$. The difference between $O(N^{1/4})$ and $O(N^{1/2})$ shows the significant improvement of these new results.

3.4.3 Error decaying property

Theorems 3.2 and 3.4 indicate that the dependence of the performance loss $\|Q^* - Q^{\pi_K}\|_{p,\rho}$ (or $\|V^* - V^{\pi_K}\|_{p,\rho_{\mathcal{X}}}$) on $(\varepsilon_k)_{k=0}^{K-1}$ is in the form of $\mathcal{E}(\varepsilon_0, \ldots, \varepsilon_{K-1}; r) = \sum_{k=0}^{K-1} \alpha_k^{2r} \|\varepsilon_k\|_{2p,\nu}^{2p}$. Since $\alpha_k \propto \gamma^{K-k}$ (3.3), this result indicates that the approximation errors at later iterations have more contribution to the final performance loss. This behavior is obscure in previous results such as Munos [2007]; Antos et al. [2008b] that the dependence of the final performance loss is expressed as $\mathcal{E}(\varepsilon_0, \ldots, \varepsilon_{K-1}; r) = \max_{k=0,\ldots,K-1} \|\varepsilon_k\|_{p,\nu}$ (see Proposition 3.1).

This property has practical and algorithmic implications too. It suggests that it is better to put more effort on having a smaller Bellman or approximation error at later iterations of API/AVI. This, for instance, can be done by gradually increasing the number of samples throughout iterations, or to use more powerful, and possibly computationally more expensive, function approximators for the later iterations of API/AVI.

To illustrate this property, we compare two different sampling schedules on a simple MDP. The MDP is a 100-state, 2-action chain similar to Chain Walk problem in the work of Lagoudakis and Parr [2003]. We use AVI with a lookup-table function representation. In the first sampling schedule, every 20 iterations we generate a fixed number of *fresh* samples by following a uniform random walk on the chain (we throw away old samples). We call this the *uniform* strategy as the samples are distributed uniformly over all iterations. In the *exponential* strategy, we again generate new samples every 20 iterations but the number of samples at the k^{th} iteration is ck^γ. The constant c is tuned such that the total number of both sampling strategies is almost the same (we give a slight margin of about 0.1% of samples in favor of the fixed strategy). What we compare is $\|Q^* - Q_k\|_{1,\nu}$ in which ν is the uniform distribution. The result is shown in Figure 3.2. The improvement of the exponential sampling schedule over the uniform one is evident as we get smaller final error when the samples are distributed according to the former strategy. Of course, one may think of more

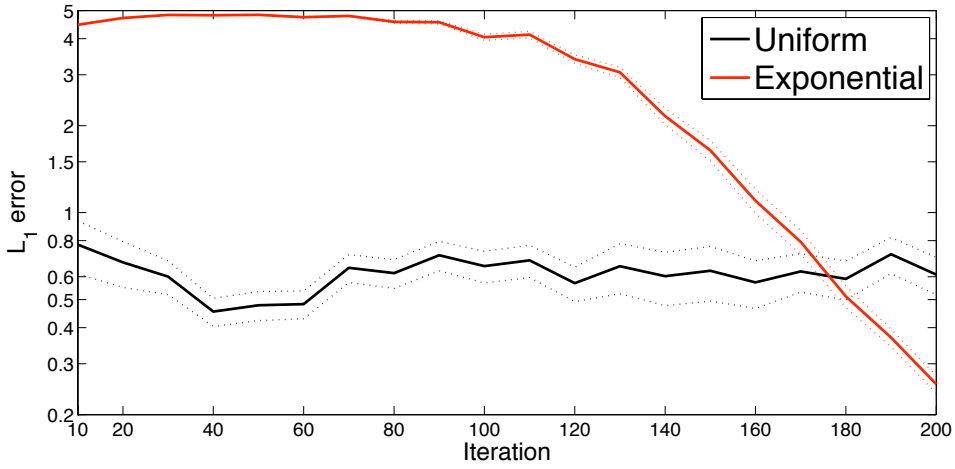

Figure 3.2: Comparison of $\|Q^* - Q_k\|_1$ for uniform and exponential data sampling schedule. The total number of samples is the same. [The Y-scale is logarithmic.]

sophisticated assignment of number of samples over iterations but this simple illustration should suffice to attract the attention of the reader to this phenomenon.

3.4.4 Restricted search over policy space

One interesting feature of our results is that it reveals more structure and restriction on the way policies can be chosen in the upper bound. To be more concrete, comparing $C_{\mathrm{PI},\rho,\nu}(K;r)$ (Theorem 3.2) and $C_{\mathrm{VI},\rho,\nu}(K;r)$ (Theorem 3.4) with $C_{\rho,\nu}$ (Proposition 3.1) indicates that:

1. Each concentrability coefficient in the definition of $C_{\mathrm{PI},\rho,\nu}(K;r)$ depends only on a single or two policies (e.g., π'_k in $c_{\mathrm{PI}_1,\rho,\nu}(K-k,m;\pi'_k)$). The same is true for $C_{\mathrm{VI},\rho,\nu}(K;r)$, and remarkably it can be seen that the policy is indeed π_K – the result of the AVI procedure. In contrast, the m^{th} term in $C_{\rho,\nu}$ has m degrees of freedom (i.e., $\pi_1,\ldots,\pi_m$), and this number is growing as $m \to \infty$.

2. The operator sup in $C_{\mathrm{PI},\rho,\nu}$ and $C_{\mathrm{VI},\rho,\nu}$ appears *outside* the summation. Therefore, we only have $K+1$ degrees of freedom (i.e., $\pi'_0,\ldots,\pi'_K$) to choose from in API and K degrees of freedom (i.e., $\pi'_1,\ldots,\pi'_K$ in AVI. On the other other hand, sup appears *inside* the summation in the definition of $C_{\rho,\nu}$. One may construct an MDP that this difference in the ordering of sup leads to an arbitrarily large ratio of two different ways of defining the concentrability coefficients.

3. In API, the definitions of concentrability coefficients $c_{\mathrm{PI}_1,\rho,\nu}$, $c_{\mathrm{PI}_2,\rho,\nu}$, and $c_{\mathrm{PI}_3,\rho,\nu}$ (Definition 3.1) imply that if $\rho = \rho^*$, the stationary distribution induced by an optimal policy π^*, then $c_{\mathrm{PI}_1,\rho,\nu}(m_1,m_2;\pi) = c_{\mathrm{PI}_1,\rho,\nu}(\cdot,m_2;\pi) = \left(\mathbb{E}\left[\left|\frac{d(\rho^*(P^\pi)^{m_2})}{d\nu}\right|^2\right]\right)^{1/2}$ with $(X,A) \sim \nu$ (similar for the other two coefficients). This special structure is hidden in the definition of $C_{\rho,\nu}$ in Proposition 3.1, and instead we have an extra m_1 degrees of flexibility.

Remark 3.3. For general MDPs, the computation of concentrability coefficients in Definition 3.1 is difficult, as it is for similar coefficients defined in Munos [2003, 2007]; Antos et al. [2008b]. The purpose of studying these coefficients is qualitative – at least for the time being.

3.5 Conclusion

To analyze an API/AVI algorithm and to study its statistical properties such as consistency or convergence rate, we require to 1) analyze the statistical properties of the algorithm running at each iteration, and 2) study the way the policy approximation/Bellman errors propagate and influence the quality of the resulting policy.

The analysis in the first step heavily uses tools from the Statistical Learning Theory (SLT) literature, e.g., Györfi et al. [2002]. In some cases, such as AVI, the problem can be cast as a standard regression with the twist that extra care should be taken to the temporal dependency of data in the RL scenario. The situation is more complicated for API procedures that directly aim for the fixed-point solution (such as LSTD and its variants). Nevertheless, still some similar tools from SLT can be used too – see Antos et al. [2008b]; Maillard et al. [2010]. We study this aspect of the analysis in Chapters 5 and 6.

The analysis for the second step is what this work has been about. In Theorems 3.2 and 3.4, we have provided upper bounds that relate the errors at each iteration of API/AVI to the performance loss of the resulting policy. These bounds are qualitatively tighter than the previous results such as those reported by Munos [2003, 2007]; Antos et al. [2008b], and provide a better understanding of what factors contribute to the difficulty of the problem. In Section 3.4, we discussed the significance of these new results and the way they improve previous ones.

Finally, we should note that there are still some unaddressed issues. Perhaps the most prominent one is to study the behavior of concentrability coefficients $c_{\mathrm{PI}_1,\rho,\nu}(m_1, m_2; \pi)$, $c_{\mathrm{PI}_2,\rho,\nu}(m_1, m_2; \pi_1, \pi_2)$, $c_{\mathrm{VI}_1,\rho,\nu}(m_1, m_2; \pi)$, and and $c_{\mathrm{VI}_2,\rho,\nu}(m_1; \pi_1, \ldots, \pi_k)$ as a function of m_1, m_2, and the transition probability kernel P of the MDP. A better understanding of them alongside a good understanding of the way each term ε_k in $\mathcal{E}(\varepsilon_0, \ldots, \varepsilon_{K-1}; r)$ behaves, help us gain more insight about the error convergence behavior of RL/Planning algorithms. The latter issue is addressed in Chapters 5 and 6.

Chapter 4

Regularized Least-Squares Regression: Learning from a β-mixing Sequence

4.1 Introduction

Our main goal in this work is to study the convergence rate of regularized least-squares regression when the covariates of the input form an exponentially β-mixing random process. Our main motivation is that the usual assumption on the independence of the input data fails to hold in a number of important practical applications. Possible relaxations of this assumption have been considered in both the statistics and machine learning communities for a long time, under assumptions of various generality. A particularly widely-used set of assumptions concerns the *mixing rate* of the input process (cf. Doukhan [1994]; Yu [1994]; Vidyasagar [2002]). [1]

The popularity of studying learning under mixing conditions is partly due to that many stochastic processes with temporal dependence are mixing. For instance, Mokkadem [1988] shows that certain ARMA processes can be modeled as an exponentially β-mixing stochastic process, the notion that we shall also use in this work. More generally, globally exponentially stable dynamical systems subjected to finite-variance continuous density input noise give rise to exponentially β-mixing Markov processes [Vidyasagar and Karandikar, 2008]. This class encompasses many dynamical systems common in the system identification and adaptive control. As the final example, the geometric ergodicity of a strictly stationary Markov chain implies exponentially (or faster) decaying β-mixing coefficients [Bradley, 2005, Theorem 3.7].

Even though some research papers consider learning in a mixing setting, only a few of them consider *regularized* empirical risk minimization. In particular, Xu and Chen [2008] study this problem in reproducing kernel Hilbert spaces (RKHS) when the input is an exponentially strongly (or, α-)mixing stationary sequence. Under an assumption similar to our metric entropy condition, they prove bounds on the estimation error. However, their bounds are suboptimal (even in the asymptotic sense), unless the input process is independent. Steinwart et al. [2009] show consistency when the squared loss is replaced by more general loss functions under relaxed conditions on the input sequence. In particular, they relax the notion of mixing and they also drop the stationarity assumption. However, they leave results concerning rates of convergence for future work. Sun and Wu [2010] replace the metric entropy condition of Xu and Chen [2008] by an assumption that requires that $L_{\mathrm{K},\mu}^{-r} m$ is square integrable with respect to the (common) distribution of the covariates μ,

[1] This chapter is the result of the collaboration of the author with Csaba Szepesvári.

where κ is the chosen kernel, $L_{\kappa,\mu}$ is the corresponding integral operator and m is the unknown regression function. Their rates, however, are not better (and sometimes worse) than those obtained by Xu and Chen [2008]. Mohri and Rostamizadeh [2010] consider a stability-based analysis. They first derive general bounds for stable algorithms for ϕ and β-mixing processes. As a corollary, they derive bounds on the estimation error for regularization empirical risk-minimization over RKHSs when the input is a ϕ-mixing stationary sequence, with the mixing coefficient decaying at a super-linear algebraic rate.

Let us now turn to the formulation of our main results. Let $\mathcal{D}_n = \big((X_1, Y_1), \dots, (X_n, Y_n)\big)$ be the input, where $X_i \in \mathcal{X}$ and $Y_i \in [-L, L]$ ($L > 0$) are random variables, and $\mathcal{X}$ is a measurable subset of a Polish space, which is a separable topological space whose topology is metrizable by a complete metric. We shall assume that $((X_t, Y_t))_{t=1,2,\dots}$ is a stationary exponentially β-mixing stochastic process (the precise definitions will given in Section 4.2.1). Let $m : \mathcal{X} \to \mathbb{R}$ be the underlying regression function $m(x) = \mathbb{E}\left[Y_i | X_i = x\right]$, and μ denote the common distribution underlying (X_i). Let

$$L(m, \hat{m}) = \int_{\mathcal{X}} |m(x) - \hat{m}(x)|^2 \mu(dx) \tag{4.1}$$

be the risk associated with the estimate $\hat{m} : \mathcal{X} \to \mathbb{R}$. Consider the regularized (or penalized) least-squares estimate $\hat{m}_n$

$$\begin{aligned} \tilde{m}_n &= \operatorname*{argmin}_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2 + \lambda_n J^2(f) \right\}, \\ \hat{m}_n(x) = T_L \tilde{m}_n(x) &= \begin{cases} L & \text{if} \quad \tilde{m}_n(x) > L, \\ \tilde{m}_n & \text{if} \quad -L \le \tilde{m}_n(x) \le L, \\ -L & \text{if} \quad \tilde{m}_n(x) < -L, \end{cases} \end{aligned} \tag{4.2}$$

where $\mathcal{F}$ is a suitable space of measurable real-valued functions with domain $\mathcal{X}$, J is the so-called regularization functional (or simply regularizer or penalizer), $\lambda_n > 0$ is the regularization coefficient, and T_L is the truncation operator.

There are various possibilities to choose the function space $\mathcal{F}$ and the regularizer J. For example, if $\mathcal{X} = (0, 1)$ and $J^2(f) = \int |f^{(k)}(x)|^2 dx$ for $k > 1$, the minimizer of (4.2) belongs to $\mathcal{F} = C^k(\mathbb{R})$, the space of k-times differentiable functions, and is in particular, will be an appropriately-defined spline function. More generally, when $\mathcal{X}$ is an open subset of $\mathbb{R}^d$, for some $k > 2d$ one may choose the regularizer $J^2(f)$ to be the sum of the squared L^2-norms of the function's k^{th} weak derivatives. In this case $\mathcal{F}$ becomes the Sobolev-space $\mathbb{W}^k(\mathbb{R}^d)$ ($= \{f : \mathcal{X} \to \mathbb{R} : J^2(f) < \infty\}$) (cf. Definition B.3 in Appendix B.1). Even more generally, one may pick $\mathcal{F}$ as an RKHS defined on domain $\mathcal{X}$ and $J^2(f) = \|f\|_{\mathcal{H}}^2$, where $\|\cdot\|_{\mathcal{H}}$ is the underlying inner-product norm of $\mathcal{F}$. Note that in all these cases (4.2) leads to a computationally tractable convex optimization problem, thanks to the representer theorem [Wahba, 1990; Schölkopf et al., 2001]. For more information about the RKHS-based approach to machine learning the reader is referred to the books by Schölkopf and Smola [2002]; Shawe-Taylor and Cristianini [2004]; Steinwart and Christmann [2008].

The main contributions of this work are as follows: First, we prove a relative deviation concentration inequality for empirical processes, generalizing Theorem 2 of Kohler [2000] from the i.i.d. processes to exponentially β-mixing, stationary stochastic processes. Next, we apply this result to the analysis of regularized least-squares regression. Under the assumptions that the true regression function belongs to the function space $\mathcal{F}$ and the input is a stationary, exponentially β-mixing sequence, and some other standard technical assumptions, we then derive a high-probability upper bound on the estimation error of this procedure. The main result shows that, e.g., for the previously mentioned Sobolev space, with an appropriate choice of the regularizer, the rate becomes the same as the optimal rate known to hold in the case when the inputs are i.i.d. random variables. The main techniques that we use are the independent-block technique [Yu, 1994; Bernstein, 1927] and the peeling

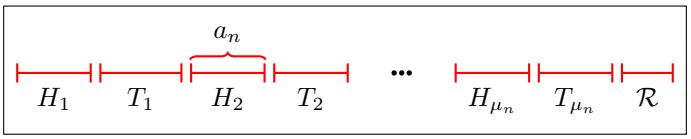

Figure 4.1: The structure of the block construction.

device [van de Geer, 2000]. To get fast rates, we have to vary the size of independent blocks according to the layer of peeling.

4.2 Definitions

The purpose of this section is to collect some definitions that we shall need later. Let $\mathbb{N}$ be the set of positive natural numbers and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. For a random variable U we shall use $\mathcal{L}(U)$ to denote its probability law. For real numbers a and b, their maximum is denoted by $a \vee b$. The norm $\|\cdot\|$ shall stand for the 2-norm of vectors.

4.2.1 Mixing Processes

In what follows, unless otherwise stated, we let $\mathcal{Z}$ denote a Polish space. Let $(Z_t)_{t=1,2,\ldots}$ be a $\mathcal{Z}$-valued stochastic process. Let $\sigma_l = \sigma(Z_1, \ldots, Z_l)$ and $\sigma'_{l+k} = \sigma(Z_{l+k}, Z_{l+k+1}, \ldots)$, where $\sigma(Z_{i_1}, Z_{i_2}, \ldots, Z_{i_k})$ is the σ-algebra for the collection $(Z_{i_1}, Z_{i_2}, \ldots, Z_{i_k})$.

Definition 4.1 (β-mixing)**.** *The k^{th} β-mixing coefficient for $(Z_t)_{t=1,2,\ldots}$ is defined as*

$$\beta_k = \sup_{l \geq 1} \mathbb{E}\left[\sup_{B \in \sigma'_{l+k}} \left| \mathbb{P}\{B|\sigma_l\} - \mathbb{P}\{B\} \right| \right].$$

The process $(Z_t)_{t=1,2,\ldots}$ is said to be β-mixing *if $\beta_k \xrightarrow{k\to\infty} 0$. Further, we say that $(Z_t)_{t=1,2,\ldots}$, is exponentially β-mixing process if for some constants $\bar{\beta}_0 \geq 0$ and $\bar{\beta}_1 > 0$, we have $\beta_k \leq \bar{\beta}_0 \exp(-\bar{\beta}_1 k)$.*

4.2.2 Independent Blocks

Fix a positive natural number $n \in \mathbb{N}$. In what follows we will need a partitioning of the set $\{1, 2, \ldots, n\}$ determined by the choice of an integral block length a_n. The partition will have $2\mu_n$ blocks with integral length a_n such that $n - 2a_n < 2\mu_n a_n \leq n$ and a "residual block":

$$\begin{aligned} H_j &= \{i : 2(j-1)a_n + 1 \leq i \leq (2j-1)a_n\}, && \text{("head")} \\ T_j &= \{i : (2j-1)a_n + 1 \leq i \leq 2ja_n\}, && \text{("tail")} \\ \mathcal{R} &= \{2\mu_n a_n + 1, \ldots, n\}, && \text{("residual")} \end{aligned}$$

for $1 \leq j \leq \mu_n$. Note that $|\mathcal{R}| < 2a_n$. Also, let $H = \cup_{1 \leq j \leq \mu_n} H_j$. See Figure 4.1 for the illustration of this construction.

Consider some sequence $(z_t)_{t=1,2,\ldots}$. We shall adopt the following conventions: For a subset S of the natural numbers $\mathbb{N}$, $\underline{z}(S)$ shall denote the ordered list $(z_i)_{i \in S}$. When S is the interval $\{i, i+1, \ldots, j\}$ for $i < j$, we shall also use $z_{i:j} = \underline{z}(S)$. Also, for $j \in \mathbb{N}$ we shall use $\underline{z}_j = (z_1, \ldots, z_j)$. These definitions are appropriately extended to the case when (z_t) is defined only for some subset of $\mathbb{N}$.

Let us introduce the *independent blocks (IB)*. Consider a $\mathcal{Z}$-valued stationary, stochastic process $(Z_t)_{t=1,2,\ldots}$. Fix n and consider $(H_j)_{1 \leq j \leq \mu_n}$ as defined above for some (a_n, μ_n). Take

a sequence of random variables $\underline{Z}'(H) = (Z'_i : i \in H)$ such that 1) $\underline{Z}'(H)$ is independent of $\underline{Z}_n$ and 2) the blocks $(\underline{Z}'(H_j) : j = 1, \dots, \mu_n)$ are independent, identically distributed and each block has the same distribution as a block from the original sequence, i.e.,

$$\mathcal{L}(\underline{Z}'(H_j)) = \mathcal{L}(\underline{Z}(H_j)) = \mathcal{L}(\underline{Z}(H_1)), \quad j = 1, \dots, \mu_n.$$

in which the second equality is because of the stationarity of the stochastic process. We refer to $\underline{Z}'(H)$ as the (μ_n, a_n)-independent block sequence underlying $\underline{Z}_n$.

The following lemma, which we shall need later, upper bounds the difference between the expectation of functions of $\underline{Z}(H)$ and $\underline{Z}'(H)$.

Lemma 4.1 (Yu [1994], Lemma 4.1)**.** *For any measurable function $h : \mathcal{Z}^{a_n \mu_n} \to \mathbb{R}$, we have*

$$\mathbb{E}\left[h(\underline{Z}(H)) - h(\underline{Z}'(H))\right] \leq \|h\|_\infty (\mu_n - 1)\beta_{a_n}\,.$$

Note that Yu only states this lemma for real-valued random variables. Since the extension to $\mathcal{Z}$-valued random variables is trivial, its proof is omitted.

4.2.3 Function Spaces

Let $\mathcal{F}$ be some space of measurable real-valued functions with a domain $\mathcal{Z}$. In order to avoid measurability problems in the case of uncountable collections of functions, throughout this work we will assume that the class $\mathcal{F}$ of functions is permissible in the sense of Pollard [1984, Appendix C]. This mild measurability condition is satisfied for most classes of functions considered in practice.

Let us now define a derived function space $\bar{\mathcal{F}}$ and some empirical norms associated to $\mathcal{F}$ and $\bar{\mathcal{F}}$. Fix n and let (a_n, μ_n) and $(H_j : 1 \leq j \leq \mu_n)$ be as in the previous section. For $f \in \mathcal{F}$, define the function $\bar{f} : \mathcal{Z}^{a_n} \to \mathbb{R}$ by

$$\bar{f}(\underline{z}_{a_n}) = \sum_{i=1}^{a_n} f(z_i),$$

and let $\bar{\mathcal{F}} = \{\bar{f} : f \in \mathcal{F}\}$. Now, fix a $\mathcal{Z}$-valued sequence $(z_t)_{t=1,2,\dots}$. We equip the spaces $\mathcal{F}$ and $\bar{\mathcal{F}}$ with the respective empirical norms $\|\cdot\|_{z_{1:n}}$ and $\|\cdot\|_{\underline{z}(H_{1:\mu_n})}$:

$$\|f\|^2_{z_{1:n}} = \frac{1}{n}\sum_{i=1}^{n} f^2(z_i)\,, \tag{4.3}$$

$$\|f\|^2_{\underline{z}(H_{1:\mu_n})} = \frac{1}{\mu_n}\sum_{j=1}^{\mu_n} \bar{f}^2(\underline{z}(H_j))\,. \tag{4.4}$$

In what follows, when $\underline{Z}_n$ is clear from the context, by a slight abuse of notation we shall use the abbreviations $\bar{f}(H_j) = \bar{f}(\underline{Z}(H_j))$ and $\bar{f}(H'_j) = \bar{f}(\underline{Z}'(H_j))$.

Let $\mathcal{M} = (\mathcal{M}, d)$ be a pseudo-metric space.[2] The *covering numbers* of a totally bounded subset B of $\mathcal{M}$ are defined for any positive $\varepsilon > 0$ as follows: The covering number $\mathcal{N}(\varepsilon, B, d)$ is the smallest number of closed d-balls of $\mathcal{M}$ that cover B. For a function space $\mathcal{G}$ with $[-M, M]$-valued functions and common domain $\mathcal{S}$, the *empirical (ℓ^2-)covering numbers* with respect to a finite sequence $s_{1:n} \in \mathcal{S}^n$ are defined as the covering numbers associated with the pseudo-metric $\|\cdot\|_{s_{1:n}}$, where this pseudo-metric is defined as in (4.3). We denote these covering numbers by $\mathcal{N}_2(\varepsilon, \mathcal{G}, s_{1:n})$. Note that this definition can be applied to both the pairs $(\mathcal{F}, \|\cdot\|_{z_{1:n}})$ and $(\bar{\mathcal{F}}, \|\cdot\|_{\underline{z}(H_{1:\mu_n})})$ and gives rise to the empirical covering numbers $\mathcal{N}_2(\varepsilon, \mathcal{F}, \|\cdot\|_{z_{1:n}})$ and $\mathcal{N}_2(\varepsilon, \bar{\mathcal{F}}, \|f\|_{\underline{z}(H_{1:\mu_n})})$. The logarithm of the covering number is called the *metric entropy*.

[2] A pseudo-metric d satisfies all properties of a metric except that $d(x, y) = 0$ does not imply that $x = y$.

4.3 Relative Deviation Concentration Inequality

In this section, we prove a general concentration inequality valid for stationary β-mixing random processes (Theorem 4.4). The result is an extension of Kohler [2000, Theorem 2] and Györfi et al. [2002, Theorem 19.3]. The proof uses the independent block technique. We start with two technical lemmas.

Lemma 4.2 (Relative Deviation Inequality)**.** *Consider a $\mathcal{Z}$-valued, stationary, β-mixing sequence $\underline{Z} = (Z_t)_{t=1,2,\ldots}$ and a permissible class $\mathcal{F}$ of real-valued functions f with domain $\mathcal{Z}$. Assume that $\sup_{f\in\mathcal{F}} \|f\|_\infty \leq M$ for some $M > 0$. Fix $n \in \mathbb{N}$ and $\varepsilon, \eta > 0$. Let $\underline{Z}'(H)$ be a (μ_n, a_n)-independent blocks sequence with a residual block $\mathcal{R}$ satisfying $\frac{|\mathcal{R}|}{n} \leq \frac{\varepsilon\eta}{6M}$. Then,*

$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}}\left|\frac{\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - \mathbb{E}\left[f(Z)\right]}{\eta + |\mathbb{E}\left[f(Z)\right]|}\right| > \varepsilon\right\} \leq 2\mathbb{P}\left\{\sup_{\bar{f}\in\bar{\mathcal{F}}}\left|\frac{\frac{1}{\mu_n}\sum_{j=1}^{\mu_n} \bar{f}(H'_j) - \mathbb{E}\left[\bar{f}(H_1)\right]}{a_n\eta + |\mathbb{E}\left[\bar{f}(H_1)\right]|}\right| > \frac{2}{3}\varepsilon\right\} + 2\beta_{a_n}\mu_n.$$

Proof. Let P denote the probability that we wish to bound. Pick any $f \in \mathcal{F}$. By the stationarity of $\underline{Z}$, the triangle inequality, and the definition of $\bar{f}$ we get

$$\begin{aligned}\left|\frac{\frac{1}{n}(\sum_{i=1}^{n} f(Z_i) - n\,\mathbb{E}\left[f(Z)\right])}{\eta + |\mathbb{E}\left[f(Z)\right]|}\right| \leq{}& \left|\frac{\frac{1}{n}(\sum_{j=1}^{\mu_n} \bar{f}(H_j) - \mu_n\mathbb{E}\left[\bar{f}(H_1)\right])}{\eta + \frac{1}{a_n}|\mathbb{E}\left[\bar{f}(H_1)\right]|}\right| \\ &+ \left|\frac{\frac{1}{n}(\sum_{j=1}^{\mu_n} \bar{f}(T_j) - \mu_n\mathbb{E}\left[\bar{f}(T_1)\right])}{\eta + \frac{1}{a_n}|\mathbb{E}\left[\bar{f}(T_1)\right]|}\right| \\ &+ \left|\frac{\frac{1}{n}(\sum_{j\in\mathcal{R}} f(Z_j) - |\mathcal{R}|\,\mathbb{E}\left[f(Z)\right])}{\eta + |\mathbb{E}\left[f(Z)\right]|}\right|.\end{aligned}$$

Since $\|f\|_\infty \leq M$, the third term is not larger than $\frac{2M|\mathcal{R}|}{\eta n}$. Now, using $\frac{|\mathcal{R}|}{n} \leq \frac{\varepsilon\eta}{6M}$ we get that this term is not larger than $\varepsilon/3$. Noting that due to the stationarity of $\underline{Z}$, the first two terms are identically distributed, so we get

$$\begin{aligned}P &\leq 2\,\mathbb{P}\left\{\sup_{\bar{f}\in\bar{\mathcal{F}}}\left|\frac{\frac{1}{n}(\sum_{j=1}^{\mu_n} \bar{f}(H_j) - \mu_n\mathbb{E}\left[\bar{f}(H_1)\right])}{\eta + \frac{1}{a_n}|\mathbb{E}\left[\bar{f}(H_1)\right]|}\right| > \frac{\varepsilon}{3}\right\} \\ &= 2\,\mathbb{P}\left\{\sup_{\bar{f}\in\bar{\mathcal{F}}}\left|\frac{\frac{2a_n}{n}(\sum_{j=1}^{\mu_n} \bar{f}(H_j) - \mu_n\mathbb{E}\left[\bar{f}(H_1)\right])}{\eta a_n + |\mathbb{E}\left[\bar{f}(H_1)\right]|}\right| > \frac{2\varepsilon}{3}\right\}.\end{aligned}$$

Since by construction $\frac{2a_n}{n} \leq \frac{1}{\mu_n}$, P can further be bounded by

$$2\mathbb{P}\left\{\sup_{\bar{f}\in\bar{\mathcal{F}}}\left|\frac{\frac{1}{\mu_n}\sum_{j=1}^{\mu_n} \bar{f}(H_j) - \mathbb{E}\left[\bar{f}(H_1)\right]}{a_n\eta + |\mathbb{E}\left[\bar{f}(H_1)\right]|}\right| > \frac{2\varepsilon}{3}\right\}.$$

Let us now apply Lemma 4.1 to bound this probability using the independent blocks sequence $\underline{Z}'(H)$. For this, choose h to be the indicator function of the event

$$\sup_{\bar{f}\in\bar{\mathcal{F}}}\left|\frac{\frac{1}{\mu_n}\sum_{j=1}^{\mu_n} \bar{f}(H_j) - \mathbb{E}\left[\bar{f}(H_1)\right]}{a_n\eta + |\mathbb{E}\left[\bar{f}(H_1)\right]|}\right| > \frac{2\varepsilon}{3}.$$

Then, $\|h\|_\infty \leq 1$. Therefore, Lemma 4.1 and $\mathcal{L}(\underline{Z}'(H_1)) = \mathcal{L}(\underline{Z}(H_1))$ gives the bound

$$P \leq 2\mathbb{P}\left\{\sup_{\bar{f}\in\bar{\mathcal{F}}}\left|\frac{\frac{1}{\mu_n}\sum_{j=1}^{\mu_n} \bar{f}(H'_j) - \mathbb{E}\left[\bar{f}(H_1)\right]}{a_n\eta + \mathbb{E}\left[\bar{f}(H_1)\right]}\right| > \frac{2\varepsilon}{3}\right\} + 2\beta_{a_n}\mu_n.$$

□

The following lemma relates the covering numbers $\mathcal{N}_2(\varepsilon, \mathcal{F}, z_{1:n})$ and $\mathcal{N}_2(\varepsilon, \bar{\mathcal{F}}, \underline{z}(H_{1:\mu_n}))$.

Lemma 4.3 (Covering Number). *For any $(z_1, \dots, z_n) \in \mathcal{Z}^n$, we have*

$$\mathcal{N}_2(\varepsilon, \bar{\mathcal{F}}, \underline{z}(H_{1:\mu_n})) \le \mathcal{N}_2\left(\tfrac{1}{2a_n}\sqrt{2(1-\tfrac{|\mathcal{R}|}{n})}\,\varepsilon, \mathcal{F}, z_{1:n}\right).$$

Proof. Pick any function $f : \mathcal{Z} \to \mathbb{R}$. Then $\|\bar{f}\|^2_{\underline{z}(H_{1:\mu_n})}$ can be bounded in terms of $\|f\|^2_{z_{1:n}}$ as follows:

$$\begin{aligned}\|\bar{f}\|^2_{\underline{z}(H_{1:\mu_n})} &= \frac{1}{\mu_n}\sum_{j=1}^{\mu_n}\left|\sum_{i\in H_j} f(z_i)\right|^2 \le \frac{a_n^2}{\mu_n a_n}\sum_{i\in H}|f(z_i)|^2 \\ &\le \frac{2a_n^2}{n(1-\frac{|\mathcal{R}|}{n})}\sum_{i=1}^{n}|f(z_i)|^2 = \frac{2a_n^2}{1-\frac{|\mathcal{R}|}{n}}\|f\|^2_{z_{1:n}}.\end{aligned}$$

Here we first applied Jensen's inequality and then we used $2a_n\mu_n = n - |\mathcal{R}|$ and that $H \subseteq \{1, \dots, n\}$.

Now consider $f_1, f_2 \in \mathcal{F}$. Using the previous inequality and $\overline{f_1 - f_2} = \bar{f}_1 - \bar{f}_2$ we get

$$\|\bar{f}_1 - \bar{f}_2\|^2_{\underline{z}(H_{1:\mu_n})} \le \frac{2a_n^2}{1-\frac{|\mathcal{R}|}{n}}\|f_1 - f_2\|^2_{z_{1:n}}.$$

Therefore any $\frac{\sqrt{2(1-\frac{|\mathcal{R}|}{n})}}{2a_n}\varepsilon$-cover of $\mathcal{F}$ is an ε-cover of $\bar{\mathcal{F}}$. □

We are ready to state the main result of this section, generalizing Theorem 2 of Kohler [2000] and Theorem 19.3 of Györfi et al. [2002] (quoted as Lemma 4.7 in the appendix) to the exponentially β-mixing stationary stochastic processes.

Theorem 4.4 (Relative Deviation Concentration Inequality). *Consider a $\mathcal{Z}$-valued, stationary, β-mixing sequence $\underline{Z} = (Z_t)_{t=1,2,\dots}$ and a permissible class $\mathcal{F}$ of real-valued functions f with domain $\mathcal{Z}$. Let $n \in \mathbb{N}$, and $K_1, K_2 \ge 1$, and choose $\eta > 0$ and $0 < \varepsilon < 1$. Assume that the following conditions hold: For any $f \in \mathcal{F}$,*

(C1) $\|f\|_\infty \le K_1$, *(uniform boundedness)*

(C2) $\mathbb{E}\left[f^2(Z)\right] \le K_2\mathbb{E}\left[f(Z)\right]$. *(variance)*

Further, consider the (a_n, μ_n)-independent blocks with the residual block $\mathcal{R}$ and assume that the following also hold:

(C3) $\sqrt{n}\varepsilon\sqrt{1-\varepsilon}\sqrt{\eta} \ge 576\left(2K_1a_n \vee \sqrt{2a_nK_2}\right)$ *(small block-size)*

(C4) $\frac{|\mathcal{R}|}{n} \le \frac{\varepsilon\eta}{6K_1}$ *and* $|\mathcal{R}| \le \frac{n}{2}$, *(small residual block)*

(C5) *For all $z_1, \dots, z_n \in \mathcal{Z}$ and all $\delta \ge \frac{\eta a_n}{8}$,*

$$\frac{\sqrt{\mu_n}\varepsilon(1-\varepsilon)\delta}{96\sqrt{2}a_n\left(K_1 \vee 2K_2\right)} \ge \int_{\frac{\varepsilon(1-\varepsilon)\delta}{16a_n(K_1\vee 2K_2)}}^{\sqrt{\delta}}\left[\log\mathcal{N}_2(\tfrac{u}{2a_n}, \mathcal{F}, z_{1:n})\right]^{\frac{1}{2}}du.$$

(small metric entropy)

Then, there exists universal constants $c_1, c_2 > 0$ such that

$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}}\left|\frac{\frac{1}{n}\sum_{i=1}^n f(Z_i) - \mathbb{E}\left[f(Z)\right]}{\eta + \mathbb{E}\left[f(Z)\right]}\right| > \varepsilon\right\} \le c_1\exp\left(-c_2\frac{\mu_n a_n \eta\,\varepsilon^2\,(1-\frac{2}{3}\varepsilon)}{a_n^2K_1^2 \vee a_nK_2}\right) + 2\beta_{a_n}\mu_n.$$

The constants can be set to $c_1 = 120$ and $c_2 = \frac{1}{2^{13}\,3^4}$.

Note that in the metric entropy condition (C5) we use the covering numbers of $\mathcal{F}$ – unlike Kohler [2000] and Györfi et al. [2002] who consider the covering numbers of a smaller subset of $\mathcal{F}$. We chose to present a simpler (weaker) result to simplify the presentation. The use of the peeling device in the proof of Theorem 4.5 obviates the need for a stronger result (Refer to Appendix B.3 for an introduction to the peeling device).

Proof. Introduce the independent blocks sequence $\{\underline{Z}'(H_j) : j = 1, \dots, \mu_n\}$ as defined in Section 4.2.2. By construction and the stationarity of the process, $\mathcal{L}(\underline{Z}'(H_j)) = \mathcal{L}(\underline{Z}(H_j)) = \mathcal{L}(\underline{Z}(H_1))$. Lemma 4.2 relates the relative deviation of the original empirical process to the relative deviation of the independent blocks process:

$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}}\left|\frac{\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - \mathbb{E}\left[f(Z)\right]}{\eta + \mathbb{E}\left[f(Z)\right]}\right| > \varepsilon\right\} \le 2\mathbb{P}\left\{\sup_{\bar{f}\in\bar{\mathcal{F}}}\left|\frac{\frac{1}{\mu_n}\sum_{j=1}^{\mu_n} \bar{f}(H'_j) - \mathbb{E}\left[\bar{f}(H_1)\right]}{a_n\eta + \mathbb{E}\left[\bar{f}(H_1)\right]}\right| > \frac{2}{3}\varepsilon\right\} + 2\beta_{a_n}\mu_n\,,$$

where we used (C1) and (C4) to verify the conditions of Lemma 4.2.

Since $(\bar{f}(H'_j))_{j=1}^{\mu_n}$ are i.i.d., we can use Theorem 19.3 of Györfi et al. [2002], which is stated as Lemma 4.7 in the appendix, to analyze the concentration of the relative deviations defined with the independent blocks by choosing n of that theorem to be the number of independent blocks μ_n and η to be $a_n\eta$. Let us now verify the conditions of this theorem:

(1) Condition (C1) implies that for any $\underline{z}_{a_n} \in \mathcal{Z}^{a_n}$ we have $|\bar{f}(\underline{z}_{a_n})| \le a_n K_1$. Let $K'_1 = a_n K_1$.

(2) Use Jensen's inequality, the stationarity of the process, and (C2) to get $\mathbb{E}\left[\bar{f}^2(H'_j)\right] = \mathbb{E}\left[\left(\sum_{i=1}^{a_n} f(Z'_i)\right)^2\right] \le a_n^2\mathbb{E}\left[f^2(Z'_1)\right] \le a_n^2 K_2\mathbb{E}\left[f(Z'_1)\right] = a_n K_2\mathbb{E}\left[\bar{f}(H'_j)\right]$. Let $K'_2 = a_n K_2$.

(3) Condition (A3) of Lemma 4.7 translates into $\sqrt{\mu_n}\varepsilon\sqrt{1-\varepsilon}\sqrt{a_n\eta} \ge 288\left(2K'_1 \vee \sqrt{2K'_2}\right)$ for $0 < \varepsilon < 1$ and $\eta > 0$. As $|\mathcal{R}| \le \frac{n}{2}$, therefore $a_n\mu_n > \frac{n}{4}$, and this condition is satisfied whenever

$$\sqrt{n}\varepsilon\sqrt{1-\varepsilon}\sqrt{\eta} \ge 576\left(2K_1 a_n \vee \sqrt{2a_n K_2}\right),$$

which is (C3).

(4) Condition (A4) of Lemma 4.7 requires that for all $\underline{z}(H_1), \dots, \underline{z}(H_{\mu_n}) \in \mathcal{Z}^{a_n}$ and all $\delta \ge \frac{a_n\eta}{8}$,

$$\frac{\sqrt{\mu_n}\varepsilon(1-\varepsilon)\delta}{96\sqrt{2}\left(K'_1 \vee 2K'_2\right)} \ge \int_{\frac{\varepsilon(1-\varepsilon)\delta}{16\left(K'_1 \vee 2K'_2\right)}}^{\sqrt{\delta}} \left[\log \mathcal{N}_2\left(u, \mathcal{B}(\bar{\mathcal{F}}, \delta), \underline{z}(H_{1:\mu_n})\right)\right]^{\frac{1}{2}} du\,, \tag{4.5}$$

where $\mathcal{B}(\bar{\mathcal{F}}, \delta) = \{\bar{f} \in \bar{\mathcal{F}} : \frac{1}{\mu_n}\sum_{j=1}^{\mu_n} \bar{f}^2(\underline{z}(H_j)) \le 16\delta\}$. Since $\mathcal{B}(\bar{\mathcal{F}}, \delta) \subset \bar{\mathcal{F}}$, we have $\mathcal{N}_2\left(u, \mathcal{B}(\bar{\mathcal{F}}, \delta), \underline{z}(H_{1:\mu_n})\right) \le \mathcal{N}_2\left(u, \bar{\mathcal{F}}, \underline{z}(H_{1:\mu_n})\right)$. According to Lemma 4.3, the latter is bounded by

$$\mathcal{N}_2(\varepsilon, \bar{\mathcal{F}}, \underline{z}(H_{1:\mu_n})) \le \mathcal{N}_2\left(\frac{1}{2a_n}\sqrt{2(1-\tfrac{|\mathcal{R}|}{n})}\,\varepsilon, \mathcal{F}, \underline{z}_{1:n}\right) \le \mathcal{N}_2\left(\frac{\varepsilon}{2a_n}, \mathcal{F}, \underline{z}_{1:n}\right).$$

Here the second inequality holds because $|\mathcal{R}| \le \frac{n}{2}$, which is satisfied by the second part of (C4). Plugging the values of K'_1 and K'_2, we get the following condition which is sufficient for (4.5):

$$\frac{\sqrt{\mu_n}\varepsilon(1-\varepsilon)\delta}{96\sqrt{2}a_n(K_1 \vee 2K_2)} \ge \int_{\frac{\varepsilon(1-\varepsilon)\delta}{16a_n(K_1 \vee 2K_2)}}^{\sqrt{\delta}} \left[\log \mathcal{N}_2\left(\frac{u}{2a_n}, \mathcal{F}, z_{1:n}\right)\right]^{\frac{1}{2}} du$$

which is in fact (C5).

Therefore the application of Lemma 4.2 and Lemma 4.7 leads to

$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}}\left|\frac{\frac{1}{n}\sum_{i=1}^{n}f(Z_i)-\mathbb{E}\left[f(Z)\right]}{\eta+\mathbb{E}\left[f(Z)\right]}\right|>\varepsilon\right\}\leq 120\exp\left(-\frac{\mu_n a_n\eta\frac{4}{9}\varepsilon^2(1-\frac{2}{3}\varepsilon)}{128\times 2304\times(a_n^2K_1^2\vee a_nK_2)}\right)+2\beta_{a_n}\mu_n\,,$$

which is the desired result. □

4.4 Analysis of Regularized Least-Squares Estimates

In this section we prove a high probability upper bound on the risk of regularized least-squares estimator (4.2) with dependent data. Theorem 4.5 shows the dependence of the error on the number of samples n and the *capacity* of the function space $\mathcal{F}$ in the asymptotic regime. The upper bound obtained is, up to a logarithmic factor, the same as the one in the i.i.d. setting.

We make the following assumptions. As before $\mathcal{X}$ is a Polish space, $\mathcal{F}$ is a permissible class of real-valued functions with domain $\mathcal{X}$. The penalty $J^2:\mathcal{F}\to\mathbb{R}$ is non-negative valued. For $R>0$, we let $\mathcal{B}_R=\{f\in\mathcal{F}:J^2(f)\leq R^2\}$.

Assumption A1 (Exponential Mixing) The process $((X_t,Y_t))_{t=1,2,\ldots}$ is an $\mathcal{X}\times\mathbb{R}$-valued, stationary, exponentially β-mixing stochastic process. In particular, the β-mixing coefficients satisfy $\beta_k\leq\bar{\beta}_0\exp(-\bar{\beta}_1 k)$, where $\bar{\beta}_0\geq 0$ and $\bar{\beta}_1>0$.

Assumption A2 (Capacity) There exist $C>0$ and $0\leq\alpha<1$ such that for any $u,R>0$ and all $x_1,\ldots,x_n\in\mathcal{X}$,

$$\log\mathcal{N}_2(u,\mathcal{B}_R,x_{1:n})\leq C\left(\frac{R}{u}\right)^{2\alpha}.$$

Assumption A3 (Boundedness) There exists $0<L<\infty$ such that the common distribution of Y_t is such that $|Y_t|\leq L$ almost surely.

Assumption A4 (Realizability) The regression function $m(x)=\mathbb{E}\left[Y_1|X_1=x\right]$ belongs to the function space $\mathcal{F}$.

Before stating the main result, we would like to remark about our assumptions.

Remark 4.1. If the mixing rate of the process is slower (e.g., $\beta_k=O(k^{-\bar{\beta}})$ for $\bar{\beta}>0$), we may still have consistent estimators that satisfy a behavior such as $\lim_{n\to\infty}\mathbb{E}\left[L(m,\hat{m}_n)\right]\to 0$ (or stronger), where $L(m,\hat{m}_n)$ is defined in (4.1). The rate of convergence, however, might be slower than what we obtain in Theorem 4.5.

Remark 4.2. The capacity Assumption A2 is mild, at least when $\mathcal{X}\subset\mathbb{R}^d$ for some $d\in\mathbb{N}$ and $\|X_t\|$ is bounded almost surely. For instance, Theorem 4 of Zhou [2003] shows its validity for a large class of RKHS with sufficiently smooth kernel functions. The reader is referred to Lemmas 20.4, 20.6 of Györfi et al. [2002], Zhou [2002, 2003], van de Geer [2000], and the discussion on pp. 226–279 of Steinwart and Christmann [2008] for some more examples.

Remark 4.3. We define the *approximation error* arising from restricting the estimators to $\mathcal{F}$ by

$$a(m;\mathcal{F})=\inf_{f\in\mathcal{F}}L(m,f)\,.$$

When $X_t\in\mathbb{R}^d$, $\|X_t\|$ and $|Y_t|$ are bounded a.s., and $\mathcal{F}$ is a Sobolev-space then $a(m,\mathcal{F})=0$ (cf. Theorem 20.4 of Györfi et al. [2002]). Therefore, a proper choice of regularization coefficient leads to a universally consistent procedure. On the other hand when $\mathcal{F}$ is "smaller", $a(m;\mathcal{F})$ might be positive. In this case let m' be the minimizer of $L(m;f)$ over $\mathcal{F}$, which we assume to exist for a moment. A simple calculation gives

$$L(m,\hat{m}_n)\leq 2\left[a(m;\mathcal{F})+L(m',\hat{m}_n)\right].$$

When the approximation error exists, the result of Theorem 4.5 can be shown to hold for the second term in the right-hand side (RHS), the so-called *estimation error*. Results regarding the behavior of the approximation error $a(m;\mathcal{F})$ for "small" RKHSs are discussed, e.g., by Smale and Zhou [2003]. Also it is notable that model selection procedures can be used to balance the estimation and approximation errors and consequently to lead to adaptive procedures with close to optimal learning rates, see e.g., Kohler et al. [2002]. The detail of the way model selection should be implemented and analyzed, however, is outside the scope of this work.

The main result of this work is as follows.

Theorem 4.5. *Let Assumptions A1–A4 hold. Define the estimate $\hat{m}_n$ by (4.2) with $\lambda_n = \left[\frac{1}{nJ^2(m)}\right]^{\frac{1}{1+\alpha}}$. There exists constants $c_1, c_2 > 0$, where c_1 depends only on L and c_2 depends only on L and $\bar{\beta}_0$, such that for any fixed $0 < \delta < 1$ and n sufficiently large,*

$$\int_{\mathcal{X}} |m(x) - \hat{m}(x)|^2 \mu(dx) \leq c_1 \left[J^2(m)\right]^{\frac{\alpha}{1+\alpha}} n^{-\frac{1}{1+\alpha}} \left[\frac{\log(n \vee c_2/\delta)}{\bar{\beta}_1}\right]^3$$

holds with probability at least $1-\delta$. In particular, when $\alpha = 0$, the above bound holds for $n \geq c_3 \exp(\bar{\beta}_1)$, while in the case of $\alpha > 0$ it holds when $n \geq c_3 \exp(\bar{\beta}_1) \vee 1/J^2(m)$ and

$$\frac{1}{n}\left(\frac{c_4 \log(n \vee c_2/\delta)}{\bar{\beta}_1}\right)^{\frac{4+5\alpha}{\alpha}} \leq J^2(m)\,, \tag{4.6}$$

where $c_3, c_4 > 0$ depends only on L.

This theorem indicates that (disregarding the logarithmic term) the asymptotic convergence rate is $O(n^{-\frac{1}{1+\alpha}})$. This is notable because it is known to be the optimal minimax rate for the i.i.d. samples under the assumption that $m \in \mathcal{F}$ and $\mathcal{F}$ has a packing entropy in the same form as in the upper bound of Assumption A2 [Yang and Barron, 1999]. Note that the choice of λ_n in the theorem depends on both α and $J(m)$, which might be unknown in practice. One can use a model selection procedure to adaptively select parameters so that the estimator achieves a rate almost as fast as the rate based on the unknown parameters of the problem. For an example of such a procedure for the i.i.d. input, refer to Kohler et al. [2002]. Let us now turn to the proof.

Proof. The proof, which is similar in spirit to that of Theorem 21.1 of Györfi et al. [2002], consists of the following main steps:

- Decompose the error into two terms $T_{1,n}$ and $T_{2,n}$ that will be defined shortly. [Step 1]
- Use the minimizer property of the empirical risk minimizer to control $T_{1,n}$. [Step 2]
- Analyze $T_{2,n}$: Apply the peeling device [Step 3], then introduce IBs that are dependent on the layer of peeling [Step 4]. Afterwards use the relative deviation concentration inequality of Theorem 4.4 to arrive at a high probability upper bound on $T_{2,n}$. [Step 5]
- Optimize the upper bound. [Step 6]

Without loss of generality in what follows we shall assume that $L \geq 1$. Let us now carry out the steps of the proof.

Step 1. Define the following error decomposition:

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

where

$$\frac{1}{2}T_{1,n} = \frac{1}{n}\sum_{i=1}^{n}\left[|\hat{m}_n(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2\right] + \lambda_n J^2(\hat{m}_n),$$
$$T_{2,n} = \mathbb{E}\left[|\hat{m}_n(X) - Y|^2|\mathcal{D}_n\right] - \mathbb{E}\left[|m(X) - Y|^2\right] - T_{1,n}.$$

Step 2. The minimizer property of $\tilde{m}_n$ and the fact that for any $u \in \mathbb{R}$, if $|Y| \le L$, then $|T_L u - Y| \le |u - Y|$ imply that

$$\begin{aligned}\frac{1}{2}T_{1,n} &\le \frac{1}{n}\sum_{i=1}^{n}\left[|\tilde{m}_n(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2\right] + \lambda_n J^2(\tilde{m}_n)\\ &\le \frac{1}{n}\sum_{i=1}^{n}\left[|m(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2\right] + \lambda_n J^2(m) = \lambda_n J^2(m).\end{aligned}$$

Therefore

$$T_{1,n} \le 2\lambda_n J^2(m). \tag{4.7}$$

Step 3. Fix any number t satisfying

$$t \ge \frac{1}{n}. \tag{4.8}$$

Our goal now is to study $\mathbb{P}\{T_{2,n} > t\}$. We have

$$\begin{aligned}\mathbb{P}\{T_{2,n} > t\} = \mathbb{P}\Bigg\{ & 2\Big(\mathbb{E}\left[|\hat{m}_n(X) - Y|^2|\mathcal{D}_n\right] - \mathbb{E}\left[|m(X) - Y|^2\right]\Big)\\ & - \frac{2}{n}\sum_{i=1}^{n}\left[|\hat{m}_n(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2\right]\\ & > t + 2\lambda_n J^2(\hat{m}_n) + \mathbb{E}\left[|\hat{m}_n(X) - Y|^2|\mathcal{D}_n\right] - \mathbb{E}\left[|m(X) - Y|^2\right]\Bigg\}.\end{aligned}$$

Let $z = (x, y)$ and define the following class of function spaces for $l = 0, 1, \cdots$:

$$G_l \triangleq \left\{g : \mathcal{X} \times \mathbb{R} \to \mathbb{R} : g(z) = |T_L f(x) - T_L y|^2 - |m(x) - T_L y|^2, f \in \mathcal{F}, J^2(f) \le \frac{2^l t}{\lambda_n}\right\}.$$

Note that functions in G_l satisfy $\|g\|_\infty \le K_1 \triangleq 4L^2$. Applying the peeling device, we get

$$\mathbb{P}\{T_{2,n} > t\} \le \sum_{l\ge 0}\mathbb{P}\left\{\sup_{g\in G_l}\frac{\mathbb{E}[g(Z)] - \frac{1}{n}\sum_{i=1}^{n} g(Z_i)}{2^l t + \mathbb{E}[g(Z)]} > \frac{1}{2}\right\}. \tag{4.9}$$

We now bound each term with the help of Theorem 4.4. For this, we shall choose an IB sequence tuned separately to each value of l.

Step 4. Fix some value of $l \in \mathbb{N}_0$. Let the block size and the number of blocks be defined by

$$a_{n,l} = \lfloor a'_{n,l}\rfloor \quad \text{and} \quad \mu_{n,l} = \left\lfloor \frac{n}{2a_{n,l}}\right\rfloor, \tag{4.10}$$

where

$$a'_{n,l} = (nt)^\gamma (2^l)^p \quad \text{and} \quad \mu'_{n,l} = \frac{n}{2a'_{n,l}} = \frac{n^{1-\gamma}}{2t^\gamma (2^l)^p}.$$

The values of $\gamma, p > 0$ will be specified later.

Note that by the assumptions $t \geq \frac{1}{n}$ and $p, \gamma > 0$, we have $a_{n,l} \geq 1$. Let $\mathcal{R}_l$ be the residual block in the $(a_{n,l}, \mu_{n,l})$-partitioning of $\{1, 2, \ldots, n\}$. The block size $a_{n,l}$, the number of blocks $\mu_{n,l}$, and the residual block size $|\mathcal{R}_l|$ have the following simple properties that will be used later:

$$n - |\mathcal{R}_l| = 2a_{n,l}\mu_{n,l} \leq n; \quad |\mathcal{R}_l| < 2a_{n,l}; \quad \mu'_{n,l} \leq \mu_{n,l}.$$

Let us show that if n and l are sufficiently large (and if γ, p satisfy certain properties) then the summands in (4.9) will be zero. We first claim that if

$$4nK_1 \leq (a'_{n,l})^{1/p} \quad \text{and} \tag{4.11}$$

$$\gamma \leq p \tag{4.12}$$

hold then $\frac{\mathbb{E}[g(Z)] - \frac{1}{n}\sum_{i=1}^n g(Z_i)}{2^l t + \mathbb{E}[g(Z)]} \leq \frac{1}{2}$. Indeed,

$$\frac{\mathbb{E}\left[g(Z)\right] - \frac{1}{n}\sum_{i=1}^n g(Z_i)}{2^l t + \mathbb{E}\left[g(Z)\right]} \leq \frac{2K_1}{2^l t}.$$

Using (4.8) and (4.12), we get $a'_{n,l} = (nt)^\gamma (2^l)^p \leq (n\,t \cdot 2^l)^p$, which is equivalent to $2^l t \geq n^{-1}(a'_{n,l})^{1/p}$. Combining this with (4.11) gives the desired statement. Now, it is easy to see that (4.11) follows from

$$p \leq \frac{1}{2} \leq 1, \tag{4.13}$$

$$a'_{n,l} \geq \frac{n}{8}, \quad \text{and} \tag{4.14}$$

$$n \geq c_1 \triangleq 4 \times 8^2 \times K_1 \geq 4^{\frac{p}{1-p}}\, 8^{\frac{1}{1-p}}\, K_1^{\frac{p}{1-p}}. \tag{4.15}$$

From now on we will assume that in addition to (4.8), the constraints (4.12), (4.13), and (4.15) hold too. Under these conditions it suffices to study the case when l is such that $a_{n,l} < n/8$.

Step 5. The following proposition, proven in the appendix, holds:

Proposition 4.6. *Consider l such that $a_{n,l} < \frac{n}{8}$. In addition, assume that*

$$0 < \gamma < p \leq \frac{1}{2+\alpha}. \tag{4.16}$$

Then, there exists constants $c_3, c_4 \geq 1$ and $c_5 > 0$, which depend only on L, such that for any

$$t > c_3^{\frac{1}{1-\gamma(2+\alpha)}} \frac{1}{n\lambda_n^{\frac{\alpha}{1-\gamma(2+\alpha)}}} + \frac{c_4}{n}, \tag{4.17}$$

we have

$$\mathbb{P}\left\{\sup_{g \in G_l} \frac{\mathbb{E}\left[g(Z)\right] - \frac{1}{n}\sum_{i=1}^n g(Z_i)}{2^l t + \mathbb{E}\left[g(Z)\right]} > \frac{1}{2}\right\} \leq 120\, \exp\left(-c_5\, \frac{\mu_{n,l}^2\, t\, 2^l}{n}\right) + 2\beta_{a_{n,l}\mu_{n,l}}.$$

We apply this proposition to the terms of the RHS of (4.9) when l is such that $a_{n,l} < n/8$. With the notation of the proposition, we get that under (4.8), (4.15), (4.16), and (4.17)

$$\begin{aligned}
\mathbb{P}\left\{T_{2,n} > t\right\} &\leq \sum_{\{l \in \mathbb{N}_0 : a_{n,l} < \frac{n}{8}\}} \left[120 \exp\left(-c_5\, \frac{\mu_{n,l}^2\, t\, 2^l}{n}\right) + 2\beta_{a_{n,l}\mu_{n,l}}\right] \\
&\leq \sum_{l \in \mathbb{N}_0} \left[120 \exp\left(-c_5\, \frac{\mu_{n,l}^2\, t\, 2^l}{n}\right) + 2\beta_{a_{n,l}\mu_{n,l}}\right].
\end{aligned}$$

Fix some $l \geq 0$. Our purpose now is to bound $\beta_{a_{n,l}}\mu_{n,l}$. By Assumption A1,

$$\beta_{a_{n,l}}\mu_{n,l} \leq \bar{\beta}_0 \exp(-\bar{\beta}_1 a_{n,l} + \log \mu_{n,l}) .$$

Thus, whenever

$$\frac{\log \mu_{n,l}}{\bar{\beta}_1 a_{n,l}} < \frac{1}{2} \tag{4.18}$$

holds, we will have $2\beta_{a_{n,l}}\mu_{n,l} \leq 2\bar{\beta}_0 \exp(-\frac{\bar{\beta}_1}{2} a_{n,l}) \leq c_6 \exp(-\frac{\bar{\beta}_1}{2} a'_{n,l})$, where $c_6 = 2\bar{\beta}_0 \exp(\frac{\bar{\beta}_1}{2})$. Using $a'_{n,l} \leq 2a_{n,l}$, $\mu_{n,l} \leq n$, and the definition of $a'_{n,l}$, we can see that (4.18) is satisfied whenever

$$t > \frac{\left(\frac{4}{\bar{\beta}_1}\log n\right)^{\frac{1}{\gamma}}}{n}. \tag{4.19}$$

Then,

$$\begin{aligned}
\mathbb{P}\left\{T_{2,n} > t\right\} &\leq \sum_{l\geq 0}\left[c_7 \exp\left(-c_5 \frac{\mu_{n,l}^2\, t\, 2^l}{n}\right) + c_6 \exp\left(-\frac{\bar{\beta}_1}{2} a'_{n,l}\right)\right] \\
&\leq \sum_{l\geq 0}\left[c_7 \exp\left(-c_8 (nt)^{1-2\gamma}(2^l)^{1-2p}\right) + c_6 \exp\left(-\frac{\bar{\beta}_1}{2}(nt)^{\gamma}(2^l)^{p}\right)\right] \\
&\leq c_9 \exp\left(-c_8 (nt)^{1-2\gamma}\right) + c_{10} \exp\left(-c_{11}\bar{\beta}_1 (nt)^{\gamma}\right) .
\end{aligned} \tag{4.20}$$

Fix some $0 < \delta < 1$. Inverting (4.20) gives that if t satisfies (4.8), (4.17) and (4.19) and if (4.15) and (4.16) hold as well then

$$T_{2,n} \leq \frac{1}{n}\left[\left(\frac{\log\left(\frac{2c_{10}}{\delta}\right)}{c_{11}\bar{\beta}_1}\right)^{\frac{1}{\gamma}} + \left(\frac{\log\left(\frac{2c_9}{\delta}\right)}{c_8}\right)^{\frac{1}{1-2\gamma}}\right]$$

holds with probability $1-\delta$.

Step 6. Combining the results of the previous steps, we find that under (4.15) and (4.16),

$$\begin{aligned}
&\int_{\mathcal{X}} |\hat{m}_n(x) - m(x)|^2 \mu(dx) = T_{1,n} + T_{2,n} \\
&\quad \leq 2\lambda_n J^2(m) + \frac{c_2^{\frac{1}{1-\gamma(2+\alpha)}}}{n\lambda_n^{\frac{\alpha}{1-\gamma(2+\alpha)}}} + \frac{(\frac{c_3}{\bar{\beta}_1}\ln\frac{c_7}{\delta})^{\frac{1}{\gamma}}}{n} + \frac{(\frac{c_4}{\bar{\beta}_1}\log n)^{\frac{1}{\gamma}}}{n} + \frac{(c_5 \ln\frac{c_7}{\delta})^{\frac{1}{1-2\gamma}}}{n} + \frac{c_6}{n}
\end{aligned} \tag{4.21}$$

holds with probability at least $1-\delta$, where we redefined the values of $c_2, \ldots, c_6, c_7 \geq 1$ in a suitable manner (Note that the values of the constants $c_2, \ldots, c_6$ depend still only on L, while c_7 depends only on L and $\bar{\beta}_0$).

Let us assume that $0 < \gamma \leq \frac{1}{3} < \frac{1}{2+\alpha}$. In this range of γ, as n gets large the third term of the RHS of (4.21) dominates the last two terms. Thus, we only need to deal with the first four terms. One can see that the choice of λ_n which minimizes the sum of these terms (disregarding the constants) is

$$\lambda_n = \left[\frac{1}{nJ^2(m)}\right]^{\frac{1-\gamma(2+\alpha)}{1-\gamma(2+\alpha)+\alpha}}, \tag{4.22}$$

which makes the sum of the first two terms proportional to

$$\lambda_n J^2(m) = \frac{[nJ^2(m)]^{\frac{\alpha}{1-\gamma(2+\alpha)+\alpha}}}{n} = \frac{e^{\frac{\alpha}{1-\gamma(2+\alpha)+\alpha}B}}{n},$$

for $B = \log(nJ^2(m))$. On the other hand, the sum of the third and fourth terms of (4.21) is upper bounded by a constant multiple of $\frac{e^{A/\gamma}}{n}$, where $A = \log(\frac{c_8}{\bar{\beta}_1} \log(c_7/\delta \vee n))$.

To choose the value of γ, we separate two cases depending on whether α is positive or zero. First, let us consider the case when $\alpha = 0$. Then, $\lambda_n J^2(m) = \frac{1}{n}$. As a result, the best choice for γ in the range $(0, \frac{1}{3}]$ is $\gamma = \frac{1}{3}$, since A/γ is decreasing in γ. Whenever $A > 0$ (i.e., $\log(c_7/\delta \vee n) \geq \bar{\beta}_1/c_8$), this choice makes the dominating term of the bound to be $e^{A/\gamma}/n = \left(\frac{c_8 \log(n \vee c_7/\delta)}{\bar{\beta}_1}\right)^3 /n$. A suitable choice for p is $p = \frac{1}{2}$. Note that $c_2^{\frac{1}{1-\gamma(2+\alpha)}} = c_2^{\frac{1}{1-\frac{2}{3}}} = c_2^3$. Whenever $n \geq 2^{10} L^2$, the constraint (4.15) is satisfied. Since the loss function is bounded, this condition can be absorbed in the constants. This finishes the proof of this case.

Consider now the case of $\alpha > 0$. The choice of γ, which unconditionally minimizes

$$\frac{1}{n}\left(e^{A/\gamma} + e^{\frac{\alpha}{1-\gamma(2+\alpha)+\alpha} B}\right)$$

is given by the solution to $A/\gamma = \frac{\alpha}{1-\gamma(2+\alpha)+\alpha} B$. Solving this for γ, we get

$$\gamma = \frac{(1+\alpha)A}{(2+\alpha)A + \alpha B}\,. \tag{4.23}$$

We will argue below that for n large enough, the chosen value satisfies $\gamma \leq \frac{1}{3}$ (and in fact $\gamma \leq \frac{1}{6}$). Thus, with this choice, the order of the terms under investigation becomes

$$\frac{1}{n} e^{A/\gamma} = \frac{1}{n} (e^B)^{\frac{\alpha}{1+\alpha}} (e^A)^{\frac{2+\alpha}{1+\alpha}} = J^2(m)^{\frac{\alpha}{1+\alpha}} n^{-\frac{1}{1+\alpha}} \left(\frac{c_8}{\bar{\beta}_1} \log(n \vee c_7/\delta)\right)^{\frac{2+\alpha}{1+\alpha}}\,. \tag{4.24}$$

Let us now show that for n large enough, we have $\gamma \leq \frac{1}{6} < \frac{1}{3}$. Indeed, as n gets large, $A = \Theta(\log\log n)$ and $B = \Theta(\log n)$. Hence, $\gamma \to 0$. In fact, a simple calculation gives that $1/6 \geq \gamma$ will be satisfied as long as n is large enough so that (4.6) holds. Moreover, $\gamma > 0$ when $A, B > 0$, which are satisfied for $n \geq \exp(\bar{\beta}_1/c_8) \vee 1/J^2(m)$. Note that any choice of p such that $0 < \gamma \leq p \leq \frac{1}{2+\alpha}$ satisfies all conditions and only affects the constants. To satisfy (4.15), it is sufficient to have $n \geq 2^{\frac{5(2+\alpha)}{1+\alpha}} L^2$. Again this condition can be absorbed in the constants. When $\gamma \leq \frac{1}{6}$, we have $\frac{1}{1-\gamma(2+\alpha)} \leq 2$. Thus, $c_2^{\frac{1}{1-\gamma(2+\alpha)}} \leq c_2^2$. This finishes the proof. □

4.5 Conclusion

Theorem 4.5 indicates that, disregarding a logarithmic factor, the rate of convergence of regularized least-squares estimates with the exponential β-mixing covariates is asymptotically the same as the minimax rate available for the i.i.d. scenario. Thus the exponential β-mixing dependence considered in this work has little effect on the efficiency of learning. It would be interesting to study this effect more closely. In particular, how far is the dependence of our bound on the rate of the β-mixing coefficients from being optimal? Another interesting issue is to design a model selection procedure with dependent inputs that achieves minimax optimal rates, e.g., along the lines of the work of Kohler et al. [2002]. For some steps towards this direction see the papers by [Meir, 2000; Modha and Masry, 1998]. Finally, it remains an interesting question of how much the dependence concepts can be relaxed while retaining the optimal minimax rates available for the i.i.d. inputs.

Appendix

4.A Proof of Proposition 4.6

In this section we prove Proposition 4.6, which was used in the proof of Theorem 4.5. For the convenience of the reader, we also quote Theorem 19.3 of Györfi et al. [2002], which is essentially the same as Theorem 2 of Kohler [2000] with some differences in constants.

Proof of Proposition 4.6. We verify the conditions of Theorem 4.4 for the choice of $\varepsilon = \frac{1}{2}$ and $\eta = 2^l t$.
(C1)–(C2): It is easy to see that these conditions are satisfied with $K_1 = 4L^2$ and $K_2 = 16L^2$ (See Györfi et al. [2002, p. 438]).
(C3): Since by assumption $L^2 \geq 1$, hence $a_{n,l} \geq 1$ implies that $2K_1 a_{n,l} > \sqrt{2a_{n,l}K_2}$. Therefore it is enough to verify that $\sqrt{n}\varepsilon\sqrt{1-\varepsilon}\sqrt{\eta} \geq 1152\, K_1 a_{n,l}$. As $a_{n,l} \leq a'_{n,l}$, it suffices to verify this condition with $a_{n,l}$ replaced by $a'_{n,l}$. Using the definition of $a'_{n,l}$, we get that (C3) is satisfied when $t \geq \frac{c'_1}{n}$ for some $c'_1 > 0$ dependent only on L.
(C4): Let us first verify $\frac{|\mathcal{R}_l|}{n} \leq \frac{\varepsilon\eta}{6K_1}$. By construction, $|\mathcal{R}_l| < 2a_{n,l} \leq 2a'_{n,l}$. Therefore, it suffices if $\frac{2a'_{n,l}}{n} < \frac{2^l t}{12K_1}$. Using the conditions on γ, p, we get that this is satisfied when $t \geq \frac{c'_2}{n}$ with some $c'_2 > 0$, dependent only on L.

Let us now verify $|\mathcal{R}_l| < \frac{n}{2}$. By assumption, we have $a_{n,l} < \frac{n}{8}$ and by construction we have $|\mathcal{R}_l| < 2a_{n,l}$, thus, $|\mathcal{R}_l| < \frac{n}{4}$.
(C5): We need to verify that for all $z_1, \ldots, z_n \in \mathcal{Z} = \mathcal{X} \times \mathbb{R}$ and all $\delta \geq \frac{2^l t\, a_{n,l}}{8}$,

$$\frac{\sqrt{\mu_{n,l}}\,\varepsilon(1-\varepsilon)\,\delta}{96\sqrt{2}\,a_{n,l}\,(K_1 \vee 2K_2)} \geq \int_{\frac{\varepsilon(1-\varepsilon)\delta}{16 a_{n,l}(K_1 \vee 2K_2)}}^{\sqrt{\delta}} \left[\log \mathcal{N}_2\left(\frac{u}{2a_{n,l}}, G_l, z_{1:n}\right)\right]^{\frac{1}{2}} du\,.$$

Let $z_t = (x_t, y_t)$, $x_t \in \mathcal{X}$, $y_t \in \mathbb{R}$. It can be shown that $\mathcal{N}_2(u, G_l, z_{1:n}) \leq \mathcal{N}_2(\frac{u}{4L}, \mathcal{F}_l, x_{1:n})$, where $\mathcal{F}_l = \left\{T_L f \in \mathcal{F} : J^2(f) \leq \frac{2^l t}{\lambda_n}\right\}$ (see Györfi et al. [2002, p. 438]). Noting that $\mu_{n,l} \geq \mu'_{n,l}$, clearly it suffices to show

$$\frac{\sqrt{\mu'_{n,l}}\,\varepsilon(1-\varepsilon)\,\delta}{96\sqrt{2}\,a_{n,l}\,(K_1 \vee 2K_2)} \geq \int_0^{\sqrt{\delta}} \left[\log \mathcal{N}_2\left(\frac{u}{8La_{n,l}}, \mathcal{F}_l, x_{1:n}\right)\right]^{\frac{1}{2}} du\,. \tag{4.25}$$

Since $\mathcal{F}_l \subset \left\{f \in \mathcal{F} : J^2(f) \leq \frac{2^l t}{\lambda_n}\right\}$, Assumption A2 indicates that

$$\mathcal{N}_2\left(\frac{u}{8La_{n,l}}, \mathcal{F}_l, x_{1:n}\right) \leq C\left(\frac{8L\, a_{n,l}\sqrt{\frac{2^l t}{\lambda_n}}}{u}\right)^{2\alpha},$$

therefore the RHS of (4.25) is upper bounded by $c'_3 a_{n,l}^{\alpha}\left(\frac{2^l t}{\lambda_n}\right)^{\frac{\alpha}{2}} \delta^{\frac{1-\alpha}{2}}$ for some constant $c'_3 > 0$, which depends only on L. Now to verify (C5), it is sufficient to prove that for $\delta \geq \frac{2^l t a_{n,l}}{8}$,

$$\frac{\sqrt{\mu'_{n,l}}\delta}{a_{n,l}} \geq c'_4 (a_{n,l})^{\alpha}\left(\frac{2^l t}{\lambda_n}\right)^{\frac{\alpha}{2}} \delta^{\frac{1-\alpha}{2}}.$$

After some manipulation we see that this condition is satisfied whenever $t \geq c'_5 \frac{a_{n,l}^{1+\alpha}}{\mu'_{n,l} 2^l \lambda_n^{\alpha}}$ for a suitably chosen $c'_5 > 0$. Using $a'_{n,l} \geq a_{n,l}$, $\mu'_{n,l} = \frac{n}{2a'_{n,l}}$, and $a'_{n,l} = (nt)^{\gamma}(2^l)^p$, we get that

it suffices to have

$$t \geq c_6' \frac{[(nt)^\gamma (2^l)^p]^{2+\alpha}}{n 2^l \lambda_n^\alpha} \iff t \geq c_7' \frac{1}{n\, \lambda_n^{\frac{\alpha}{1-\gamma(2+\alpha)}} (2^l)^{\frac{1-p(2+\alpha)}{1-\gamma(2+\alpha)}}},$$

where $c_7' = (c_6')^{\frac{1}{1-\gamma(2+\alpha)}}$ and we used the assumption that $\gamma < \frac{1}{2+\alpha}$. For $\gamma < p \leq \frac{1}{2+\alpha}$, the value of $(2^l)^{\frac{1-p(2+\alpha)}{1-\gamma(2+\alpha)}}$ is a non-decreasing function of l, so the metric entropy condition (C5) is satisfied if

$$t \geq c_7' \frac{1}{n \lambda_n^{\frac{\alpha}{1-\gamma(2+\alpha)}}}.$$

By taking $c_3 = c_6'$ and $c_4 = c_1' \vee c_2'$, all the conditions of the Theorem 4.4 are satisfied. Therefore,

$$\mathbb{P}\left\{ \sup_{g \in G_l} \frac{\mathbb{E}\left[g(Z)\right] - \frac{1}{n}\sum_{i=1}^n g(Z_i)}{2^l t + \mathbb{E}\left[g(Z)\right]} > \frac{1}{2} \right\} \leq 120 \exp\left(- \frac{\mu_{n,l}^2 \left(2^l t\right) \left(\frac{1}{2}\right)^2 \left(1 - \frac{2}{3}\cdot\frac{1}{2}\right)}{9 \times 32 \times 1152 \left(4L^2\right)^2 n} \right) + 2\beta_{a_n} \mu_n \,.$$

which we benefitted from the fact that for $L \geq 1$, we have $a_{n,l}^2 K_1^2 \geq a_{n,l} K_2$ in addition to $a_{n,l}\mu_{n,l} \leq \frac{n}{2}$. This is the desired result after absorbing all constants into $c_5 > 0$. □

Lemma 4.7 (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 $K_1, K_2 \geq 1$, $0 < \varepsilon < 1$, $\eta > 0$, 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 $\delta \geq \eta/8$,*

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

Then,

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

Chapter 5

Regularized Fitted Q-Iteration Algorithm

5.1 Introduction

In this chapter, we introduce *Regularized Fitted Q-Iteration (RFQI)* – a regularization-based AVI approach to solve RL/Planning problems with large state spaces. This chapter's contributions are two-fold: algorithmic and theoretical.[1]

The algorithmic contribution of this work is to formulate the RFQI method as an AVI procedure that uses regularized least-squares regression at each iteration (Section 5.2). There we develop a generic RFQI algorithm and provide a closed-form solution when the estimated value function is chosen from the class of *reproducing kernel Hilbert spaces (RKHS)* [Wahba, 1990]. This kernel-based formulation is intriguing because RKHSs are general, flexible, and easy to incorporate prior knowledge [Shawe-Taylor and Cristianini, 2004].

The theoretical contribution of this chapter is to analyze the statistical properties of RFQI (Section 5.3) and to provide an upper bound on the quality of the resulting policy and its relation to the performance of the optimal policy (Theorem 5.8 in Section 5.3.5). We show how the performance depends on the number of samples, the capacity of the function space to which the estimated value function belongs, and some intrinsic properties of the MDP.

The analysis of RFQI has four main steps. First we ask how large the performance loss of the resulting policy will be if the sizes of errors of each iteration are known. We answer this question in Section 5.3.1 by using the result of Chapter 3. We then focus on a single iteration of RFQI and study the statistical behavior of the corresponding regularized least-squares regression with dependent input covariates. Part of the analysis is done in Section 5.3.2 in which the material is mostly borrowed from Chapter 4 (also Farahmand and Szepesvári [2012]). We observe that the upper bound on the error of each iteration depends on the number of samples, the capacity of the function space to which the estimator belongs, the smoothness of the target regression function, and the function approximation error of representing the target in the function space. As opposed to the conventional supervised learning scenarios, the function approximation error and the smoothness depend on the result of previous iterations as if they "propagate" throughout iterations. We study these phenomena, which are specific to AVI/RFQI, in Section 5.3.3 (Behavior of the Function Approximation Error) and Section 5.3.4 (Behavior of the Smoothness). All these lead to Theorem 5.8 that reveals some aspects of learning a close-to-optimal policy that have not been known beforehand in the work of Munos and Szepesvári [2008], which analyzes Fitted

[1] This chapter is the result of the collaboration of the author with Csaba Szepesvári, Mohammad Ghavamzadeh, and Shie Mannor.

Q-Iteration algorithm and is the closest theoretical result to this work. Theorem 5.8 indicates that upon the proper choice of parameters, the dependence of the sample complexity for the task of estimating the optimal value function on the capacity of the function space is minimax optimal. Also it is seen that the size of the function approximation error and the smoothness of the target function at each iteration depend on the results of the previous iterations. We discuss various aspects of the main result in Section 5.4.

In Section 5.5 we briefly discuss the l_1-regularization-based formulation of RFQI, and finally in Section 5.6 we summarize the chapter and suggest some topics for future study.

5.2 Algorithm

RFQI is an approximate value iteration method that belongs to the class of Fitted Q-Iteration algorithms [Ernst et al., 2005; Riedmiller, 2005; Munos and Szepesvári, 2008]. In this section, we first describe the RFQI algorithm in terms of a sequence of optimization problems and afterwards we show how these optimization problems may be solved when the function space is an RKHS.

The RFQI algorithm (Algorithm 1) works as follows: It receives the number of iterations K, the initial action-value function Q_0, a collection of K datasets $\mathcal{D}^{(0)}, \mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(K-1)}$, with the cardinalities of $m_0, m_1, \ldots, m_{K-1}$, respectively, a set of function spaces $\mathcal{F}_0^{|\mathcal{A}|}, \cdots, \mathcal{F}_{K-1}^{|\mathcal{A}|}$ and their corresponding regularizers $J_k : \mathcal{F}^{|\mathcal{A}|} \to \mathbb{R}$ $(k = 0, 1, \ldots, K-1)$, and a set of regularization coefficients $\lambda_0, \cdots, \lambda_{K-1}$. In the setup that we shall later analyze, we assume that for $k = 0, \cdots, K-1$,

$$\mathcal{D}^{(k)} = \big((X_1^{(k)}, A_1^{(k)}, R_1^{(k)}, X_1'^{(k)}), \ldots, (X_{m_k}^{(k)}, A_{m_k}^{(k)}, R_{m_k}^{(k)}, X_{m_k}'^{(k)})\big),$$

all satisfying the offline sampling assumption of Section 2.2.1. We also denote the collection of all datasets as $\mathcal{D}_n = \big(\mathcal{D}^{(0)}, \mathcal{D}^{(1)}, \cdots, \mathcal{D}^{(K-1)}\big)$ with $n = m_0 + \ldots + m_{K-1}$. To simplify the analysis, we assume that $\mathcal{D}^{(k)}$ and $\mathcal{D}^{(l)}$ $(k \neq l)$ are independent datasets. In practice, however, one may use $\mathcal{D}^{(0)} = \ldots = \mathcal{D}^{(K-1)}$, which corresponds to reusing all the available data at each iteration. An analysis that handles this scenario has been done in Munos and Szepesvári [2008].

The RFQI algorithms starts from an initial action-value function Q_0. At iteration k, it approximately performs a single step of value iteration and finds a Q_{k+1} that is close to T^*Q_k, i.e., $Q_{k+1} \approx T^*Q_k$. This is done by solving the following regularized least-squares regression problem:

$$Q_{k+1} = \operatorname*{argmin}_{Q \in \mathcal{F}_k^{|\mathcal{A}|}} \Big\| Q(X_i, A_i) - \hat{T}^* Q_k(X_i, A_i) \Big\|^2_{\mathcal{D}^{(k)}} + \lambda_k J_k^2(Q), \tag{5.1}$$

where $\|\cdot\|^2_{\mathcal{D}^{(k)}}$ is the empirical norm defined in Section 2.2.1, and $\hat{T}^*Q_k$ is the single-sample empirical estimate of T^*Q_k, which is also based on $\mathcal{D}^{(k)}$. Here $\mathcal{F}_k^{|\mathcal{A}|}$ is the action-value function space, $J_k(Q)$ is the corresponding nonnegative-valued regularizer (or penalizer) that penalizes the "roughness" of Q, and $\lambda_k > 0$ is the regularization coefficient. We call the value of $J_k(Q)$ the smoothness of Q, even though it might not coincide with the conventional derivative-based notions of smoothness.

To see the connection of this algorithm to value iteration, note that the first term in the RHS of (5.1) is the sample-based squared error of using $Q(X_i, A_i)$ to predict $R(X_i, A_i) + \gamma \max_{a' \in \mathcal{A}} Q_k(X_i', a')$ at $(X_i, A_i) = (X_i^{(k)}, A_i^{(k)})$ for $i = 1, \ldots, m_k$. This term is the empirical estimate of the loss

$$L_k(Q) = \mathbb{E}\left[|Q(X, A) - T^*Q_k(X, A)|^2 \Big| Q_k \right]$$

with $(X, A) \sim \nu$. Fitting an action-value function Q that minimizes this L_2-loss corresponds to the regression problem where the covariates are $(X_i, A_i) \in \mathcal{X} \times \mathcal{A}$ and the regression

Algorithm 1 RFQI(K;Q_0;$\{\mathcal{D}^{(k)}\}_{k=0}^{K-1}$; $\{\mathcal{F}_k^{|\mathcal{A}|}\}_{k=0}^{K-1}$; $\{J_k\}_{k=0}^{K-1}$; $\{\lambda_k\}_{k=0}^{K-1}$)

// K: Number of iterations
// Q_0: Initial action-value function
// $\mathcal{D}^{(0)}, \ldots, \mathcal{D}^{(K-1)}$: Batch of samples for each iteration
// $\mathcal{F}_0^{|\mathcal{A}|}, \cdots, \mathcal{F}_{K-1}^{|\mathcal{A}|}$: The action-value function spaces
// $J_0, \cdots, J_{K-1}$: The regularizers
// $\lambda_0, \cdots, \lambda_{K-1}$: The regularization coefficients
for $k = 0$ to $K-1$ **do**
  $Q_{k+1} \leftarrow \operatorname{argmin}_{Q \in \mathcal{F}_k^{|\mathcal{A}|}} \left\| Q(X_i, A_i) - \hat{T}^* Q_k(X_i, A_i) \right\|_{\mathcal{D}^{(k)}}^2 + \lambda_k J_k^2(Q)$
end for
return Q_K and $\hat{\pi}(\cdot; Q_K)$

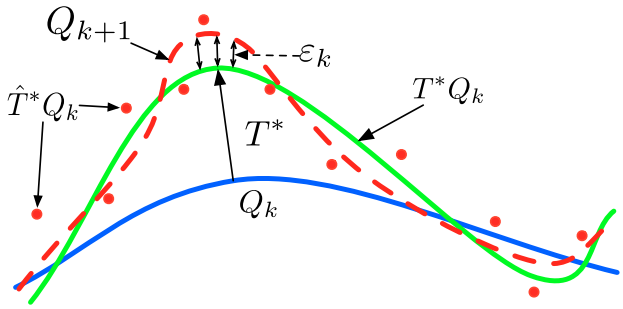

Figure 5.1: Fitted Q-Iteration Procedure. Given Q_k (blue curve), it generates $\hat{T}^* Q_k$ (red dots), and fits an action-value function Q_{k+1} (red dashed curve). The difference between Q_{k+1} and $T^* Q_k$ (green curve) is the function ε_k.

function is

$$\mathbb{E}\left[r(x,a) + \gamma \max_{a' \in \mathcal{A}} Q_k(X', a') \,\middle|\, Q_k, X = x, A = a \right] = (T^* Q_k)(x,a),$$

which is indeed the target of the exact value iteration algorithm $Q_{k+1} = T^* Q_k$. The RFQI algorithm solves this regression problem with the use of the regularized least-squares regression estimator.

As a consequence of the finite sample size and the function approximation error, Q_{k+1} is not equal to $T^* Q_k$ and there will be a residual error $\varepsilon_k = T^* Q_k - Q_{k+1}$. As the performance of any AVI algorithm, including RFQI, critically depends on $\|\varepsilon_k\|$ $(k = 0, \ldots, K-1)$ (Section 5.3.1), it is desirable for the estimation algorithm to make sure these error are as small as possible given a limited number of samples. Regularized least-squares regression is an example of a sample-efficient method that upon the proper choice of $\mathcal{F}_k^{|\mathcal{A}|}$, J_k, and λ_k can do this task.

Among possible choice for $\mathcal{F}^{|\mathcal{A}|}$, such as finite-dimensional linear spaces, infinite-dimensional function spaces defined by growing neural networks or decision trees, and Sobolev spaces, we focus on the case when $\mathcal{F}^{|\mathcal{A}|}$ is an RKHS [Wahba, 1990; Steinwart and Christmann, 2008]. Let $\mathcal{F}_k^{|\mathcal{A}|} = \mathcal{H}_k$ be the RKHS used in the k^{th} iteration. The natural choice for the regularizer is the inner-product norm of the RKHs itself, that is $J(Q) = \|Q\|_{\mathcal{H}}$. The RKHS formulation of (5.1) then becomes

$$Q_{k+1} = \operatorname*{argmin}_{Q \in \mathcal{F}_k^{|\mathcal{A}|}[=\mathcal{H}_k]} \left\| Q(X_i, A_i) - \hat{T}^* Q_k(X_i, A_i) \right\|_{\mathcal{D}^{(k)}}^2 + \lambda_k \|Q\|_{\mathcal{H}_k}^2 . \tag{5.2}$$

Even though solving (5.1) for general function spaces might be difficult, it becomes computationally tractable when we pick an RKHS and formulate the problem as (5.2). If $\mathrm{K}_k : (\mathcal{X} \times \mathcal{A}) \times (\mathcal{X} \times \mathcal{A}) \to \mathbb{R}$ denotes the unique reproducing kernel underlying $\mathcal{H}_k$, by the application of the Representer Theorem for RKHSs (e.g., Schölkopf et al. [2001, Theorem 4.2] quoted as Theorem B.1 in Appendix B.1.1) we get that every solution to (5.2) is the sum of kernels centered on the observed samples:

$$Q_{k+1}(x,a) = \sum_{i=1}^{m_k} \alpha_i^{(k+1)} \, \mathrm{K}_k \left((X_i^{(k)}, A_i^{(k)}), (x,a) \right),$$

where $\boldsymbol{\alpha}^{(k+1)} = (\alpha_1, \ldots, \alpha_{m_k})^\top$ are the coefficient that should be determined. Let us assume that Q_k was also obtained in a similar form:

$$Q_k(x,a) = \sum_{i=1}^{m_{k-1}} \alpha_i^{(k)} \, \mathrm{K}_{k-1} \left((X_i^{(k-1)}, A_i^{(k-1)}), (x,a) \right).$$

Replacing Q in (5.2) by its expansion and using the fact that $\|Q\|_{\mathcal{H}_k}^2 = \boldsymbol{\alpha}^\top \boldsymbol{K}_k \boldsymbol{\alpha}$ with $\boldsymbol{K}_k$ being the Grammian matrix, to be specified shortly, we get

$$\boldsymbol{\alpha}^{(k+1)} = \operatorname*{argmin}_{\boldsymbol{\alpha} \in \mathbb{R}^{m_k}} \left\| \boldsymbol{r}^{(k)} + \gamma \boldsymbol{K}_k^+ \boldsymbol{\alpha}^{(k)} - \boldsymbol{K}_k \boldsymbol{\alpha} \right\|_{\mathcal{D}^{(k)}}^2 + \lambda_k \, \boldsymbol{\alpha}^\top \boldsymbol{K}_k \boldsymbol{\alpha}, \tag{5.3}$$

with $\boldsymbol{K}_k \in \mathbb{R}^{m_k \times m_k}$ and $\boldsymbol{K}_k^+ \in \mathbb{R}^{m_k \times m_{k-1}}$ defined as

$$[\boldsymbol{K}]_{ij} = \mathrm{K}_k \left((X_i^{(k)}, A_i^{(k)}), (X_j^{(k)}, A_j^{(k)}) \right),$$
$$[\boldsymbol{K}^+]_{ij} = \mathrm{K}_{k-1} \left((X_i'^{(k)}, A_i^{*(k)}), (X_j^{(k-1)}, A_j^{(k-1)}) \right),$$

where $A_j^{*(k)} = \operatorname{argmax}_{a \in \mathcal{A}} Q_k(X_j'^{(k)}, a)$, and $\boldsymbol{r}^{(k)} = (R_1^{(k)}, \ldots, R_k^{(k)})^\top$. Solving (5.3) for $\boldsymbol{\alpha}$, we obtain the following closed-form solution:

$$\boldsymbol{\alpha}^{(k+1)} = \begin{cases} (\boldsymbol{K}_0 + m_0 \lambda_0 \boldsymbol{I})^{-1} \left(\boldsymbol{r}^{(0)} + \gamma \mathbf{Q_0} \right) & k = 0, \\ (\boldsymbol{K}_k + m_k \lambda_k \boldsymbol{I})^{-1} \left(\boldsymbol{r}^{(k)} + \gamma \boldsymbol{K}_k^+ \boldsymbol{\alpha}^{(k)} \right) & k = 1, \ldots, K-1, \end{cases}$$

in which $\mathbf{Q_0} = \left(Q_0(X_1'^{(0)}, A_1^{*(0)}), \ldots, Q_0(X_{m_0}'^{(0)}, A_{m_0}^{*\,(0)}) \right)^\top$.

The computational complexity of the k^{th} iteration with a naive implementation is $O(m_k^3)$ as it uses matrix inversion. Therefore if we divide the total number of samples n to K equal-sized chunks, the computational cost of K iterations of RFQI is $O(\frac{n^3}{K^2})$.

Remark 5.1. The above algorithm can also be used in a parametric setting, which might be preferable to the nonparametric approach if one has significant prior knowledge about the action-value function. Let $\Phi^{[p]}(\cdot,\cdot) : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^p$ be the feature vector defined by p basis functions. Define the function space $\mathcal{F}^{|\mathcal{A}|}(p) = \{\Phi^{[p]}(\cdot,\cdot)^\top \theta | \theta \in \mathbb{R}^p\}$. For any choice of positive semi-definite matrix Λ, one can define the l_2-norm of $Q(\cdot,\cdot;\theta) \in \mathcal{F}^{|\mathcal{A}|}(p)$ as $J^2(Q) = \theta^\top \Lambda \theta$. Notice that the estimation problem at each iteration is the conventional *ridge* regression estimator [Hoerl and Kennard, 1970]. The other possibility is to use the l_1-norm as the regularizer. This is discussed in Section 5.5.

Remark 5.2. In this work, we present and analyze the RFQI algorithm when the task is to find an approximate optimal action-value function. Nevertheless, RFQI can also be modified to evaluate a given policy π by changing $\hat{T}^*$ to $\hat{T}^\pi$ in Algorithm 1. We do not analyze this modification in this work.

Remark 5.3. Regularization has a Bayesian interpretation too. The L_2-regularized least-squares regression estimator is equivalent to finding the maximum a posteriori estimate in

the Gaussian Process regression framework with a Gaussian prior over the space of functions; and the l_1-regularization is equivalent to having a Laplacian prior over the space of functions, see Rasmussen and Williams [2006, Section 6.2]. We do not however follow Bayesian approach to derive our results, mainly because proving consistency/convergence bounds for the posteriors can be problematic.

5.3 Theoretical Analysis

The goal of the analysis is to provide an upper bound on the performance loss of policy π_K returned by RFQI, as measured by $\|Q^* - Q^{\pi_K}\|_{1,\rho}$. The probability measure ρ is chosen by the user to specify the relative importance of various regions of the state-action space, which in general is different from the sample distribution ν.

The analysis has four steps. In the first step we use the result of Chapter 3 to study how the fitting errors $\|Q_{k+1} - T^*Q_k\|_\nu$ $(k = 0, \dots, K-1)$ propagate throughout iterations and affect the performance loss of π_K (Section 5.3.1). The remaining three steps are concerned with bounding $\|Q_{k+1} - T^*Q_k\|_\nu$. The starting point, presented in Section 5.3.2, is an error bound available for regularized least-squares regression with β-mixing inputs – borrowed from Chapter 4. The presented error bound consists of two terms, one bounding the *approximation error* and the other bounding the *estimation error*. The approximation error measures the loss of using the best approximation to T^*Q_k from $\mathcal{F}_k^{|\mathcal{A}|}$, while the estimation error bounds the random variation of Q_{k+1}, which arises because the procedure uses a finite random sample. The estimation error at iteration k mainly depends on the number of samples, the capacity of the function space $\mathcal{F}_k^{|\mathcal{A}|}$, and $J(T^*Q_k)$. Because of the iterative nature of RFQI procedure, the analysis is more complicated than the conventional supervised learning scenario. The difference is that as opposed to the supervised learning scenario, which has a fixed target function, we deal with a random target function that depends on the result of previous iterations. This affects both the function approximation error and the smoothness of the target function as if they propagate throughout iterations. We study the behavior of the approximation error and the smoothness in Sections 5.3.3 and 5.3.4, respectively. Combining these results, we obtain the main theorem of this chapter (Theorem 5.8).

We note that throughout our analysis, the regularization coefficients λ_ks are chosen in such a manner so that the resulting bounds are minimized. As such, λ_k would depend on unknown quantities. The reason this should not be of major concern is because one can use data-dependent model-selection methods to tune λ_k and still achieve essentially the same performance, see e.g., [Bartlett et al., 2002; Arlot and Celisse, 2009].

5.3.1 Error Propagation for Approximate Value Iteration

Let $Q_0, Q_1, \dots, Q_K \in \mathcal{F}^{|\mathcal{A}|}$ be a sequence of action-value functions, perhaps generated by some approximate value iteration procedure that approximates T^*Q_k by Q_{k+1}.[2] Let the error at iteration k be

$$\varepsilon_k = T^*Q_k - Q_{k+1}. \tag{5.4}$$

Further, let π_K be the policy greedy w.r.t. Q_K and $p \geq 1$. In this section, we use Theorem 3.4 of Section 3.3 to relate the performance loss $\|Q^* - Q^{\pi_K}\|_{p,\rho}$ to the ν-weighted L_{2p}-norms of the error sequence $(\varepsilon_k)_{k=0}^{K-1}$. This performance loss indicates the regret of following the policy π_K instead of an optimal policy when the initial state-action is distributed according to ρ. To relate these two measures that are entangled through the MDP, we define the following concentrability coefficients.

[2] To enhance the flow of the reading, Section 3.3 is partly repeated here.

Definition 5.1 (Expected Concentrability of the Future State-Action Distribution)**.** *Given $\rho, \nu \in \mathcal{M}(\mathcal{X} \times \mathcal{A})$, $m \geq 0$, and an arbitrary sequence of stationary policies $(\pi_m)_{m \geq 1}$, let $\rho P^{\pi_1} P^{\pi_2} \dots P^{\pi_m} \in \mathcal{M}(\mathcal{X} \times \mathcal{A})$ denote the future state-action distribution obtained after m transitions, when the first state-action pair is distributed according to ρ and then we follow the sequence of policies $(\pi_k)_{k=1}^m$. For integers $m_1, m_2 \geq 1$, policy π and the sequence of policies $\pi_1, \dots, \pi_k$ define the concentrability coefficients*

$$c_{VI_1,\rho,\nu}(m_1, m_2; \pi) \triangleq \left(\mathbb{E}\left[\left| \frac{d\left(\rho (P^{\pi})^{m_1} (P^{\pi^*})^{m_2}\right)}{d\nu}(X, A) \right|^2 \right] \right)^{\frac{1}{2}},$$

$$c_{VI_2,\rho,\nu}(m_1; \pi_1, \dots, \pi_k) \triangleq \left(\mathbb{E}\left[\left| \frac{d\left(\rho (P^{\pi_k})^{m_1} P^{\pi_{k-1}} P^{\pi_{k-2}} \dots P^{\pi_1}\right)}{d\nu}(X, A) \right|^2 \right] \right)^{\frac{1}{2}},$$

where $(X, A) \sim \nu$. If the future state-action distribution $\rho (P^{\pi})^{m_1} (P^{\pi^*})^{m_2}$ (similarly, $\rho (P^{\pi_k})^{m_1} P^{\pi_{k-1}} P^{\pi_{k-2}} \dots P^{\pi_1}$) is not absolutely continuous w.r.t. ν, we let $c_{VI_1,\rho,\nu}(m_1, m_2; \pi) = \infty$ (similarly, $c_{VI_2,\rho,\nu}(m_1; \pi_1, \dots, \pi_k) = \infty$).

The concentrability coefficients are used in a change of measure argument. Due to the dynamics of MDP and AVI, this change depends not only on ν and ρ, but also on the transition kernels P^{π} and P^{π^*}, see e.g., Munos [2007] and Chapter 3 of this thesis. In order to compactly present our results, we define

$$a_k = \frac{(1-\gamma)\gamma^{K-k-1}}{1-\gamma^{K+1}}, \quad 0 \leq k < K, \tag{5.5}$$

and for $0 \leq r \leq 1$,

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

$$C_{\mathrm{VI},\rho,\nu}(K; r) =$$

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

where in the last definition the supremum is taken over all policies.

Theorem 5.1 (Error Propagation for AVI – Theorem 3.4 in Section 3.3)**.** *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 $(Q_k)_{k=0}^K \subset B(\mathcal{X} \times \mathcal{A}, Q_{max})$, and the corresponding sequence $(\varepsilon_k)_{k=0}^{K-1}$ defined in (5.4), we have*

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

We discuss the significance of this result and compare it to the previous work such as Munos [2007] in Section 5.4.2.

5.3.2 Error Bounds for Regularized Regression

The goal of this section is to analyze the statistical behavior of the fitting procedure that leads to the error ε_k at iteration k. The main result of this section, Theorem 5.2, relates $\left\| \varepsilon_k \right\|_{\nu}$ to the sample size m_k, the capacity of the function space $\mathcal{F}_k^{|\mathcal{A}|}$, and the intrinsic difficulty of the problem, characterized by the smoothness of the (random) target function $T^* Q_k$. We start by listing our assumptions required for the result of this section.

Assumption A5 (Function Space) The subset $\mathcal{F}^{|\mathcal{A}|} \subset B(\mathcal{X} \times \mathcal{A})$ is a separable and complete Carathéodory set.

In addition to the usual measurability requirement, in order to avoid the measurability issues caused by taking supremum over an uncountable function space $\mathcal{F}^{|\mathcal{A}|}$, we require the space to be a separable and complete Carathéodory set in the sense defined in Section 7.3 of Steinwart and Christmann [2008] (quoted in Appendix B.4).

Assumption A6 (Regularizer) For all values of $0 \leq k \leq K-1$, define two regularizer functionals $J_k : B(\mathcal{X}) \to \mathbb{R}$ and $J_k : B(\mathcal{X} \times \mathcal{A}) \to \mathbb{R}$ that are pseudo-norms on $\mathcal{F}_k$ and $\mathcal{F}_k^{|\mathcal{A}|}$, respectively.[3] For all $Q \in \mathcal{F}_k^{|\mathcal{A}|}$ and $a \in \mathcal{A}$, we have $J_k(Q(\cdot, a)) \leq J_k(Q)$.

Note that commonly-used regularizers are pseudo-norms. In particular, RKHS norms are in fact norms, and thus are also pseudo-norms. 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 $J'^2(Q) = \sum_{a \in \mathcal{A}} \|Q(\cdot, a)\|_{\mathcal{H}}^2$ for $\mathcal{H}$ being the RKHS defined on $\mathcal{X}$.

Assumption A7 (Function Space Capacity) Let $a \in \mathcal{A}$, $R > 0$, $0 \leq k \leq K-1$ and define the "ball" $\mathcal{B}_{k,a,R} = \{Q(\cdot, a) \in \mathcal{F}_k : J_k^2(Q(\cdot, a)) \leq R^2\}$. There exists constants $C > 0$ and $0 \leq \alpha < 1$ such that for all $k = 0, \ldots, K-1$, $a \in \mathcal{A}$, $u, R > 0$, and $x_1, \ldots, x_n \in \mathcal{X}$, the following "metric entropy condition" [Györfi et al., 2002; van de Geer, 2000] holds:

$$\log \mathcal{N}_2(u, \mathcal{B}_{k,a,R}, x_{1:n}) \leq C \left(\frac{R}{u}\right)^{2\alpha}.$$

This is a standard assumption, which is satisfied by a large number of function spaces of interest, including Sobolev spaces and various RKHS. Refer to van de Geer [2000]; Zhou [2002, 2003] and Steinwart and Christmann [2008] for several examples. An alternative assumption would be to have a similar metric entropy for the balls in $\mathcal{F}_k^{|\mathcal{A}|}$ (instead of $\mathcal{F}_k$). This would slightly change a few steps of the proofs, but leave the results essentially the same. Moreover, it makes the second part of Assumption A6 (that is $J_k(Q(\cdot, a)) \leq J_k(Q)$) unnecessary. Nevertheless, as results on the capacity of $\mathcal{F}$ is more common in the statistical learning literature, we stick to the combination of Assumptions A6 and A7. For the convenience of the reader, the definition of the metric entropy is provided in Appendix B.2.

Assumption A8 (Sampling) For all values of $0 \leq k \leq K-1$, the stochastic process $\left((X_1^{(k)}, A_1^{(k)}), \ldots, (X_{m_k}^{(k)}, A_{m_k}^{(k)})\right)$ is an $\mathcal{X} \times \mathbb{R}$-valued strictly stationary, exponentially β-mixing process with marginal ν. The β-mixing coefficients satisfy $\beta_k \leq \bar{\beta}_0 \exp(-\bar{\beta}_1 k)$, where $\bar{\beta}_0 \geq 0$ and $\bar{\beta}_1 > 0$. Furthermore, $X_t'^{(k)} \sim P(\cdot | X_t^{(k)}, A_t^{(k)})$ for $t = 1, \ldots, m_k$.

Even though many stochastic processes of interest are exponentially β-mixing (see Chapter 4), one can still consider a slower mixing (e.g., $\beta_k = k^{-\bar{\beta}}$ for $\bar{\beta} > 0$) at the price of obtaining slower convergence rates.

Assumption A9 (Boundedness) There exists $0 < Q_{\max} < \infty$ such that the common distribution of $\hat{T}^* Q(X_t, A_t)$ satisfies $|\hat{T}^* Q(X_t, A_t)| \leq Q_{\max}$ almost surely.

If an *a priori* bound on the immediate expected rewards (or the random rewards) is known, this assumption can always be enforced by possibly truncating the estimates.

[3] Note that here we are slightly abusing notation 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 a specific expression the identity of the regularizer should always be clear from the context.

Assumption A10 (Independence of Data Sets) $\mathcal{D}^{(k)}$ and $\mathcal{D}^{(l)}$ are independent for $k \neq l$.

We rely on Assumption A10 to simplify the proof. One may follow the same line of argument as Munos and Szepesvári [2008] to handle the scenario that $\mathcal{D}^{(0)} = \mathcal{D}^{(1)} = \cdots = \mathcal{D}^{(K-1)}$.

The next theorem is the main result of this section and provides an upper bound on $\|\varepsilon_k\|_\nu$. It is a slightly modified form of the result proven in Chapter 4. The first difference concerns that here we need to deal with the simultaneous estimation of $|\mathcal{A}|$ functions (i.e., action-value functions), while the result that we build on concerned only the estimation of a single function. For further details on this difference, see Appendix 5.A. The second difference is that the statement of this theorem allows the regularization coefficient to be larger than the optimal choice. The validity of this new statement can easily be concluded from the original proof.

In this section and later, we assume that the regularization coefficients are chosen according to

$$\lambda_k = B \left[\frac{1}{m_k J^2(\Pi_{\nu,\mathcal{F}_k^{|\mathcal{A}|}} T^* Q_k)} \right]^{\frac{1}{1+\alpha_k}}, \tag{5.6}$$

where $\Pi_{\nu,\mathcal{F}_k^{|\mathcal{A}|}}$ is the projection operator defined in Section 2.2.1, and $B \geq 1$.

Theorem 5.2. *[Regularized Regression for Mixing Processes – Theorem 4.5 in Chapter 4] Let Assumptions A5–A10 hold. Define the estimate Q_{k+1} by (5.1) with the choice of λ_k as in (5.6). There exists constants $c_k, c'_k > 0$, where c_k depends only on Q_{max} and c'_k depends only on Q_{max} and $\bar{\beta}_0$, such that for any fixed $0 < \delta < 1$ and m_k sufficiently large,*

$$\|Q_{k+1} - T^* Q_k\|_\nu^2 \leq \inf_{Q' \in \mathcal{F}_k^{|\mathcal{A}|}} \|Q' - T^* Q_k\|_\nu^2 + B\, c_k \left[J^2(\Pi_{\nu,\mathcal{F}_k^{|\mathcal{A}|}} T^* Q_k) \right]^{\frac{\alpha_k}{1+\alpha_k}} m_k^{-\frac{1}{1+\alpha_k}} \left[\frac{\log(m_k \vee c'_k/\delta)}{\bar{\beta}_1} \right]^3,$$

holds with probability at least $1-\delta$. In particular, when $\alpha = 0$, the above bound holds for $m_k \geq c''_k \exp(\bar{\beta}_1)$, while in the case of $\alpha > 0$ it holds when $m_k \geq c''_k \exp(\bar{\beta}_1) \vee 1/J^2(\Pi_{\nu,\mathcal{F}_k^{|\mathcal{A}|}} T^* Q_k)$ and

$$\frac{1}{m_k} \left(\frac{c'''_k \log(m_k \vee c'_k/\delta)}{\bar{\beta}_1} \right)^{\frac{4+5\alpha}{\alpha}} \leq J^2(\Pi_{\nu,\mathcal{F}_k^{|\mathcal{A}|}} T^* Q_k), \tag{5.7}$$

where $c''_k, c'''_k > 0$ depends only on Q_{max}.

The first term of the bound, $\inf_{Q' \in \mathcal{F}^{|\mathcal{A}|}} \|Q' - T^* Q_k\|_\nu^2$, defines the function approximation error, while the second term is called the estimation error. Notice that both the function approximation error and the smoothness term $J^2(\Pi_{\nu,\mathcal{F}_k^{|\mathcal{A}|}} T^* Q_k)$ are random. The analysis of the behavior of these terms will be the subject of the next two sections.

Remark 5.4. If the function space $\mathcal{F}_k^{|\mathcal{A}|}$ is rich enough (e.g., a universal kernel is used), the function approximation for the class of continuous functions shall be zero. On the contrary, 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]. The question of how to optimally balance the estimation and the function approximation errors by the choice of function space $\mathcal{F}_k^{|\mathcal{A}|}$ is beyond the scope of this work, but can be formulated as a model selection problem, see e.g., Smale and Zhou [2003]; Steinwart and Christmann [2008].

5.3.3 The Behavior of the Function Approximation Error

The goal of this section is to study the behavior of the function approximation error, $\inf_{Q' \in \mathcal{F}_k^{|\mathcal{A}|}} \|Q' - T^* Q_k\|_\nu$. Previously, Munos and Szepesvári [2008] bounded this error by the so-called *inherent Bellman error*,

$$a(T^* \mathcal{F}_{k-1}^{|\mathcal{A}|}; \mathcal{F}_k^{|\mathcal{A}|}) = \sup_{Q \in \mathcal{F}_{k-1}^{|\mathcal{A}|}} \inf_{Q' \in \mathcal{F}_k^{|\mathcal{A}|}} \|Q' - T^* Q\|_\nu , \tag{5.8}$$

which gives a deterministic, *a priori* upper bound on the error, though this bound can be very conservative. The reason is because Q_k can be expected to reside in a small vicinity of $(T^*)^k Q_0$. In this case, the actual function approximation error is expected to be close to $\inf_{Q' \in \mathcal{F}_k^{|\mathcal{A}|}} \|Q_{k+1} - (T^*)^k Q_0\|_\nu$, which might be much smaller than $a(T^* \mathcal{F}_{k-1}^{|\mathcal{A}|}; \mathcal{F}_k^{|\mathcal{A}|})$. The purpose of this section is to formalize this intuition.

We need the following, new concentrability coefficients, which are similar to those introduced earlier in Section 5.3.1.

Definition 5.2 (Concentrability Coefficient of One-step Transitions)**.** *Let ν be a distribution over the state-action pairs, $(X, A) \sim \nu$, $\nu_{\mathcal{X}}$ the marginal distribution of X, and $\pi_b(\cdot|\cdot)$ the conditional probability of A given X. Further, let P be a transition probability kernel $P : \mathcal{X} \times \mathcal{A} \to \mathcal{M}(X)$ and $P_{x,a} = P(\cdot|x, a)$. Define the concentrability coefficient of one-step transitions w.r.t. ν by*

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

where $C_{AE}(\nu; P) = \infty$ if $P_{x,a}$ is not absolutely continuous w.r.t. $\nu_{\mathcal{X}}$ for some $(x, a) \in \mathcal{X} \times \mathcal{A}$, or if $\pi_b(a'|y) = 0$ for some $(y, a') \in \mathcal{X} \times \mathcal{A}$.

The constant $C_{\mathrm{AE}}(\nu; P)$ is large if after one step of transition, the future state can be highly concentrated at some state where the probability of taking some action a' is small as well as $\nu_{\mathcal{X}}$. Hence, the name "concentrability of one-step transitions".

The main result of this section is the theorem stated below. The proof is provided in Appendix 5.B.

Theorem 5.3. *Let $(Q_k)_{k=0}^{K-1}$ be a sequence of state-action value functions, $b_k = \|T^* Q_k - Q_{k+1}\|_\nu$, $0 \le k \le K-1$. Then, it holds for any $0 \le k \le K-1$ that*

$$\inf_{Q' \in \mathcal{F}_k^{|\mathcal{A}|}} \|Q' - T^* Q_k\|_\nu \le \inf_{Q' \in \mathcal{F}_k^{|\mathcal{A}|}} \left\| Q' - (T^*)^{(k+1)} Q_0 \right\|_\nu + \sum_{i=0}^{k-1} \left(\gamma \, C_{AE}(\nu; P) \right)^{i+1} b_{k-1-i}.$$

Note that the above bound will be smaller than the *a priori* bound (5.8) when the individual errors and the one-step concentrability coefficient are all small enough. In the limit of a large number of examples, the error bounds b_k will be shown to be arbitrarily close to zero, hence, in the limit the bound in this theorem is better, assuming finite one-step concentrability. Of course, under the assumptions of the theorem, both the bound of this theorem and (5.8) hold at the same time, therefore one can always take the smallest of the two bounds.

5.3.4 The Behavior of the Smoothness

In this section, we study the behavior of the smoothness term $J_k(\Pi_{\mathcal{F}_k^{|\mathcal{A}|}} T^* Q_k)$. Our analysis has two steps. With the help of an assumption on the MDP, we relate $J_k(\Pi_{\mathcal{F}_k^{|\mathcal{A}|}} T^* Q_k)$ to $J_{k-1}(Q_k)$. We also upper bound $J_{k-1}(Q_k)$ in terms of $J_{k-1}(\Pi_{\mathcal{F}_{k-1}^{|\mathcal{A}|}} T^* Q_{k-1})$. This gives rise to a recursive bound.

To upper bound $J_k(Q_{k+1})$, we use Theorem 6.8 of Chapter 6, which itself is an extension of Theorem 10.2 of van de Geer [2000].[4] Note that this result is stated for i.i.d. data, so it only covers the planning scenario. Nevertheless, we expect the same result to hold true for exponentially β-mixing inputs too, but showing this is left for future work.

Proposition 5.4. *Fix $0 \le k \le K-1$. Consider the regularized regression problem defined in (5.1) with the regularization coefficient λ_k chosen according to (5.6). Let Assumptions A5, A6, A7, A9, and A10 hold, and in addition, assume that the sequence $\left((X_1^{(k)}, A_1^{(k)}), \dots, (X_{m_k}^{(k)}, A_{m_k}^{(k)})\right)$ is i.i.d. Then, there exists a constant $c_k > 0$ such that for any $m_k \in \mathbb{N}$ and $0 < \delta < 1$, we have*

$$J_k(Q_{k+1}) \le c_k \, J_k(\Pi_{\nu, \mathcal{F}_k^{|\mathcal{A}|}} T^* Q_k) \sqrt{\ln(1/\delta_k)},$$

with probability at least $1 - \delta_k$.

Proposition 5.4 relates the smoothness of Q_{k+1} (the result of the k^{th} iteration of RFQI) to the smoothness of $J_k(\Pi_{\mathcal{F}_k^{|\mathcal{A}|}} T^* Q_k)$. We may recursively relate this smoothness to Q_{k-1}, $Q_{k-2}, \dots, Q_0$ if 1) the operator $\Pi_{\mathcal{F}_k^{|\mathcal{A}|}} T^*$ is well-behaving in the sense that $J_k(\Pi_{\mathcal{F}_k^{|\mathcal{A}|}} T^* Q_k)$ is not much different from $J_k(Q_k)$ and 2) the regularizer J_{k-1} penalizes similar to J_k. This is required because we want to avoid a situation that a smooth function in the function space $\mathcal{F}_{k-1}$ is rough in the function space $\mathcal{F}_k$. To formalize these requirements, we make the following assumption.

Assumption A11 For any $k = 0, 1, \dots, K-1$ and for any $Q \in \mathcal{F}_k^{|\mathcal{A}|}$, there exist some constants $0 \le L_R, L_P < \infty$, depending only on the MDP and the set of $\{\mathcal{F}_i^{|\mathcal{A}|}\}_{i=0}^{K-1}$ such that

$$J_k(\Pi_{\nu, \mathcal{F}_k^{|\mathcal{A}|}} T^* Q) \le L_R + \gamma L_P J_{k-1}(Q).$$

where for the notational convenience, we set $J_{-1} \triangleq J_0$.

The validity of this assumption for a certain class of MDPs is shown in Proposition 6.16 in Chapter 6.

The main result of this section is an immediate corollary of Proposition 5.4 and Assumption A11. Since the previous proposition is not proved for β-mixing inputs, we state the conclusion of the previous proposition as a condition of the next result.

Proposition 5.5. *Let Assumption A11 hold. Pick $0 < \delta \le 1$ and assume that for some $c > 0$, $J_k(Q_{k+1}) \le c \, J_k(\Pi_{\nu, \mathcal{F}_k^{|\mathcal{A}|}} T^* Q_k) \sqrt{\ln(1/\delta)}$ holds for $k = 0, \dots, K-1$. Then, for any $0 \le k \le K-1$, it holds that*

$$J_k(\Pi_{\mathcal{F}_k^{|\mathcal{A}|}} T^* Q_k) \le L_R + \gamma L_P \left[J_0(Q_0) \left[L'_P(\delta)\right]^k + \frac{L'_R(\delta)}{1 - L'_P(\delta)} \left(1 - \left[L'_P(\delta)\right]^k\right) \mathbb{I}\{k \ge 1\} \right],$$

where $L'_R(\delta) = c \, L_R \sqrt{\ln(1/\delta)}$ and $L'_P(\delta) = c \gamma L_P \sqrt{\ln(1/\delta)}$.

Proof. The proof is a simple induction. For completeness, it is given in Appendix 5.C. □

In order to simplify the final upper bound, we would like L'_P to be smaller than 1. Of course, this may not hold true in many cases including when $\delta \ll 1$. The crucial point, however, is that c_k defined in Proposition 5.4, and as a result $L'_P(\delta)$, may be changed by the change of λ_k defined in (5.6). Essentially what we require is to oversmooth the estimate at all iterations. The following assumption ensures that one can oversmooth without too much increasing of the regularization coefficient.

[4] Actually, the result quoted here is less general than the aforementioned theorem. That theorem holds uniformly on any target function, but here we fix the target. As a result, we only have $J_k(\Pi_{\nu, \mathcal{F}_k^{|\mathcal{A}|}} T^* Q_k)$ instead of $J_k(\Pi_{\nu, \mathcal{F}_k^{|\mathcal{A}|}} T^* Q_k) + J_k(Q_k)$ in the upper bound.

Assumption A12 Consider the regularized regression problem defined in (5.1) with the choice of $\lambda_k = \lambda$ and denote $J_k(Q_{k+1}(\cdot;\lambda))$ as a function of λ by $J_k(\lambda)$. For a given value of $J \in \mathbb{R}$, let $J_k^{-1}(J)$ be the value of λ such that $J_k(\lambda) = J$. For any $0 < \rho \leq 1$ and $\lambda \leq \lambda_0 < \infty$, there exists a finite positive constant Λ, such that for all $k = 0, 1, \ldots, K-1$, we have

$$\frac{J_k^{-1}(\rho J_k(\lambda))}{\lambda} \leq \frac{\Lambda}{\rho}.$$

5.3.5 Main Result

In this section, we derive a high probability error upper bound for the performance loss of the RFQI algorithm based on the results of previous sections. Proposition 5.6 upper bounds $\|Q_{k+1} - T^* Q_k\|_\nu$. Proposition 5.7 simplifies the bound when Assumption A12 holds. Finally Theorem 5.8, which is the main result of this work, upper bounds the performance loss $\|Q^* - Q^{\pi_K}\|_{1,\rho}$.

Fix $K \in \mathbb{N}$, let $0 < \delta < 1$ be a fixed constant, and pick $\delta' = \frac{\delta}{2K}$. Let $L'_R(\delta') = c\, L_R \sqrt{\ln(1/\delta')}$ and $L'_P(\delta') = c\, \gamma L_P \sqrt{\ln(1/\delta')}$ with c defined as in Proposition 5.5. Let $1 \leq B < \infty$ and for $k = 0, 1, \cdots, K-1$, choose the regularization coefficient according to (5.6). With these choices of $(\lambda_k)_{k=0}^{K-1}$, let the sequence $(Q_{k+1})_{k=0}^{K-1}$ be defined as the solution of (5.1). For any real-valued sequence $b_0, \cdots, b_{k-1}$, define $c_{\mathrm{A}}(b_0, b_1, \cdots, b_{k-1})$ and $c_{\mathrm{E}}^{(1)}(k; L_R, L_P)$ as

$$c_{\mathrm{A}}(b_0, b_1, \cdots, b_{k-1}) = 2\gamma \left(\sum_{i=1}^{k} \left(\gamma C_{\mathrm{AE}}^2(\nu; P) \right)^i \right) \left(\sum_{i=0}^{k-1} \gamma^i\, b_{k-1-i}^2 \right), \tag{5.9}$$

$$c_{\mathrm{E}}^{(1)}(k; L_R, L_P) = B\, c'_k(Q_{\max}) \times \left[L_R + \gamma L_P \left(\frac{L'_R(\delta')}{1 - L'_P(\delta')} + \left(J_0(Q_0) - \frac{L'_R(\delta')}{1 - L'_P(\delta')} \right) [L'_P(\delta')]^k \right) \right]^{\frac{2\alpha_k}{1+\alpha_k}},$$

for a constant $c'_k(Q_{\max})$, which is a function of $Q_{\max}$ and $\bar{\beta}_0$ only. Now define $(b_i(\delta))_{i=0}^{k-1}$ as

$$b_k^2(\delta) = c_{\mathrm{E}}^{(1)}(k; L_R, L_P)\, m_k^{-\frac{1}{1+\alpha_k}} \left[\frac{\log(m_k \vee \frac{K}{\delta})}{\bar{\beta}_1} \right]^3 + 2 \inf_{Q' \in \mathcal{F}_k^{|\mathcal{A}|}} \left\| Q' - (T^*)^{(k+1)} Q_0 \right\|_\nu^2 + c_{\mathrm{A}}(b_0(\delta), \cdots, b_{k-1}(\delta)). \tag{5.10}$$

Proposition 5.6. *Let assumptions A5, A6, A7, A9, A10, and A11 hold. Also assume that either 1) Assumption A8 holds* ***and*** *Proposition 5.4 holds for the β-mixing processes (learning scenario), or 2) Assumption A8 is strengthened to hold only for the i.i.d. processes (planning scenario). Fix $0 < \delta < 1$ and let b_k be defined according to (5.10). Then for $m_0, \ldots, m_k$ sufficiently large and for any $0 \leq k \leq K-1$ we have*

$$\|Q_{k+1} - T^* Q_k\|_\nu^2 \leq b_k^2(\delta)$$

with probability at least $1 - \frac{k}{K}\delta$.

For small δ' and fixed B, the value of $L'_P(\delta')$ might be large. This results in $c_{\mathrm{E}}^{(1)}(k; L_R, L_P)$ being an exponentially growing function of k. Nevertheless, when Assumption A12 holds, we show that if one picks $B(\delta') = \Theta(\sqrt{\ln(1/\delta')}) \vee 1$, the value of $L'_P(\delta')$ of the new estimate would be smaller than one. This is stated in the next proposition. Let

$$c_{\mathrm{E}}^{(2)}(k; L_R, L_P) = c''_k(Q_{\max}) L_P \left[L_R + \gamma L_P J_0(Q_0) \right]^{\frac{2\alpha_k}{1+\alpha_k}},$$

for a constant $c_k''(Q_{\max}) > 0$, which is a function of $Q_{\max}$ and $\bar{\beta}_0$ only. Define $(b_i(\delta))_{i=0}^{k-1}$ as

$$b_k^2(\delta) = c_{\mathrm{E}}^{(2)}(k; L_R, L_P)\, m_k^{-\frac{1}{1+\alpha_k}} \left[\frac{\log(m_k \vee \frac{K}{\delta})}{\bar{\beta}_1}\right]^{7/2} + 2 \inf_{Q' \in \mathcal{F}_k^{|\mathcal{A}|}} \left\| Q' - (T^*)^{(k+1)} Q_0 \right\|_\nu^2 + c_{\mathrm{A}}(b_0(\delta), \cdots, b_{k-1}(\delta)). \tag{5.11}$$

Proposition 5.7. *Let all assumptions of Proposition 5.6 hold. In addition, let Assumption A12 hold. Fix $0 < \delta < 1$ and let b_k be defined according to (5.11). There exists a constant $c(L_P, \gamma)$ such that if $B(\delta') = c(\gamma) L_P \sqrt{\ln(1/\delta')} \vee 1$, and $m_0, \ldots, m_k$ are sufficiently large, for any $0 \le k \le K-1$ we have*

$$\|Q_{k+1} - T^* Q_k\|_\nu^2 \le b_k^2(\delta),$$

with probability at least $1 - \frac{k}{K}\delta$.

The next theorem, which is the main result of this work, upper bounds the performance loss $\|Q^* - Q^{\pi_K}\|_{1,\rho}$.

Theorem 5.8. *Let the assumptions of Proposition 5.6 or Proposition 5.7 hold. Choose a fixed $0 < \delta < 1$ and let $(b_k(\delta))_{k=0}^{K-1}$ be defined as (5.10) (when the assumptions of Proposition 5.6 hold) or (5.11) (when the assumptions of Proposition 5.7 hold). Assume that $C_{VI,\rho,\nu}$ is finite. Define a_k according to (5.5). Let*

$$\mathcal{E}(b_0(\delta), \ldots, b_{K-1}(\delta); r) = \sum_{k=0}^{K-1} a_k^{2r}\, b_k^2(\delta). \tag{5.12}$$

Then, the ρ-weighted performance loss of π_K is upper bounded by

$$\|Q^* - Q^{\pi_K}\|_{1,\rho} \le \frac{2\gamma}{(1-\gamma)^2} \left[\inf_{r \in [0,1]} C_{VI,\rho,\nu}^{\frac{1}{2}}(K; r) \mathcal{E}^{\frac{1}{2}}(b_0(\delta), \ldots, b_{K-1}(\delta); r) + \frac{2}{1-\gamma}\gamma^K R_{max}\right], \tag{5.13}$$

with probability at least $1 - \delta$.

Proof of Proposition 5.6, Proposition 5.7, and Theorem 5.8. Fix $\delta' > 0$. For $i = 0, \ldots, K-1$, invoke Theorem 5.2 to provide an upper bound for the ν-weighted L_2-norm of the ε_i defined in (5.4). There exist constants c_i and $c_{1,i}$ such that for sufficiently large m_i (so that the conditions on m_i in Theorem 5.2 are satisfied), we have

$$\|Q_{i+1} - T^* Q_i\|_\nu^2 \le b_i'^2,$$
$$b_i' \triangleq \inf_{Q' \in \mathcal{F}_i^{|\mathcal{A}|}} \|Q' - T^* Q_i\|_\nu^2 + B\, c_i(Q_{\max}) \left[J^2(\Pi_{\nu, \mathcal{F}_i^{|\mathcal{A}|}} T^* Q_i)\right]^{\frac{\alpha_i}{1+\alpha_i}} m_i^{-\frac{1}{1+\alpha_i}} \left[\frac{\log(m_i \vee c_{1,i}/\delta')}{\bar{\beta}_1}\right]^3, \tag{5.14}$$

with probability at least $1 - \delta'$. Consider the event $\mathcal{E}_k^{(1)}$ such that $\|Q_{i+1} - T^* Q_i\|_\nu \le b_i'$ holds for all $0 \le i \le k-1$. The probability of this event is at least $1 - k\delta'$.

Furthermore, according to Proposition 5.4, there exists a constant $c_i > 0$, independent of m_i and δ', such that

$$J_i(Q_{i+1}) \le c_i\, J_i(\Pi_{\nu, \mathcal{F}_i^{|\mathcal{A}|}} T^* Q_i) \sqrt{\ln(1/\delta')}, \tag{5.15}$$

with probability at least $1 - \delta'$. Consider the event $\mathcal{E}_k^{(2)}$ such that (5.15) holds for $0 \le i \le k-1$. The probability of this event is at least $1 - k\delta'$. Consider the event $\mathcal{E}_k = \mathcal{E}_k^{(1)} \cap \mathcal{E}_k^{(2)}$.

The probability of the event $\mathcal{E}_k$ is at least $1-2k\delta'$. From now on, our analysis will be on the event $\mathcal{E}_k$.

We use techniques developed in Sections 5.3.3 and 5.3.4 to control the function approximation error $\inf_{Q'\in\mathcal{F}_k^{|\mathcal{A}|}}\|Q'-T^*Q_k\|_\nu$ and the smoothness $J(\Pi_{\nu,\mathcal{F}_k^{|\mathcal{A}|}}T^*Q_k)$, respectively. Theorem 5.3 and the Cauchy-Schwarz inequality imply that

$$\begin{aligned}\inf_{Q'\in\mathcal{F}_k^{|\mathcal{A}|}}\|Q'-T^*Q_k\|_\nu^2 \le& 2\inf_{Q'\in\mathcal{F}_k^{|\mathcal{A}|}}\left\|Q'-(T^*)^{(k+1)}Q_0\right\|_\nu^2\\ &+2\left(\sum_{i=1}^{k}\left(\gamma\,C_{\mathrm{AE}}^2(\nu;P)\right)^i\right)\left(\gamma\sum_{i=0}^{k-1}\gamma^i\,b_{k-1-i}'^2\right).\end{aligned}\tag{5.16}$$

To bound the smoothness term $J(\Pi_{\nu,\mathcal{F}_k^{|\mathcal{A}|}}T^*Q_k)$ in (5.14), we use Proposition 5.5. On the event $\mathcal{E}_k^{(2)}\subset\mathcal{E}_k$ and with our fixed choice of δ' for all $i=0,\dots,k-1$, the conditions of the proposition are satisfied and therefore for $k\ge 0$,

$$J_k(\Pi_{\nu,\mathcal{F}_k^{|\mathcal{A}|}}T^*Q_k)\le L_R+\gamma L_P\left[\left(J_0(Q_0)-\frac{L_R'(\delta')}{1-L_P'(\delta')}\right)[L_P'(\delta')]^k+\frac{L_R'(\delta')}{1-L_P'(\delta')}\right].\tag{5.17}$$

Note that in (5.16), we may replace $(b_i')_{i=0}^{k-1}$, which are random, with any deterministic upper bounds and the inequality still holds. With this in mind, (5.14) alongside (5.16) and (5.17) lead to

$$\begin{aligned}&\|Q_{k+1}-T^*Q_k\|_\nu^2\le b_k^2\\ &\triangleq 2\inf_{Q'\in\mathcal{F}_k^{|\mathcal{A}|}}\left\|Q'-(T^*)^{(k+1)}Q_0\right\|_\nu^2+2\gamma\left(\sum_{i=1}^{k}\left(\gamma\,C_{\mathrm{AE}}^2(\nu;P)\right)^i\right)\left(\sum_{i=0}^{k-1}\gamma^i\,b_{k-1-i}^2\right)\\ &+B\,c_k(Q_{\max})\left[L_R+\gamma L_P\left(\frac{L_R'(\delta')}{1-L_P'(\delta')}+\left(J_0(Q_0)-\frac{L_R'(\delta')}{1-L_P'(\delta')}\right)[L_P'(\delta')]^k\right)\right]^{\frac{2\alpha_k}{1+\alpha_k}}\times\\ &\quad m_k^{-\frac{1}{1+\alpha_k}}\left[\frac{\log(m_k\vee c_{1,k}/\delta')}{\bar\beta_1}\right]^3,\end{aligned}$$

on the event $\mathcal{E}_k$. One may absorb $\log^3(2c_{1,k})$ into c_k to have a new constant c_k'. Noting that $\mathbb{P}\{\mathcal{E}_k\}\ge 1-2k\delta'=1-\frac{k}{K}\delta$ finishes the proof of Proposition 5.6.

To simplify the bound, we use Assumption A12 with the choice of $\rho=1/(2L_P'(\delta'))\wedge 1/2=\Theta(1/\sqrt{\ln(1/\delta')})\wedge 1/2$. This ensures that if $B(\delta')=\Gamma/\rho=2(L_P'(\delta')\wedge 1)\Gamma\le 2c\gamma\Gamma L_P\sqrt{\ln(1/\delta')}=\Theta(L_P\sqrt{\ln(1/\delta')})$, then (5.17) holds with $L_R''(\delta')\le L_R'(\delta')/(2L_P'(\delta'))=L_R/(2\gamma L_P)$ and $L_P''(\delta')\le 1/2$ replacing $L_R'(\delta')$ and $L_P'(\delta')$, respectively. This simplifies (5.17) to

$$J_k(\Pi_{\nu,\mathcal{F}_k^{|\mathcal{A}|}}T^*Q_k)\le L_R+\gamma\frac{L_PL_R''(\delta')}{1-1/2}+\gamma L_PJ_0(Q_0)(1/2)^k\le 2L_R+\gamma L_PJ_0(Q_0).$$

This leads to a simplified upper bound

$$\begin{aligned}\|Q_{k+1}-T^*Q_k\|_\nu^2\le b_k^2\triangleq& 2\inf_{Q'\in\mathcal{F}_k^{|\mathcal{A}|}}\left\|Q'-(T^*)^{(k+1)}Q_0\right\|_\nu^2\\ &+2\gamma\left(\sum_{i=1}^{k}\left(\gamma\,C_{\mathrm{AE}}^2(\nu;P)\right)^i\right)\left(\sum_{i=0}^{k-1}\gamma^i\,b_{k-1-i}^2\right)\\ &+c_k''(Q_{\max})L_P\left[L_R+\gamma L_PJ_0(Q_0)\right]^{\frac{2\alpha_k}{1+\alpha_k}}m_k^{-\frac{1}{1+\alpha_k}}\left[\frac{\log(m_k\vee\frac{K}{\delta})}{\bar\beta_1}\right]^{7/2}.\end{aligned}$$

As before, $\mathbb{P}\{\mathcal{E}_k\} \geq 1 - \frac{k}{K}\delta$. This finishes the proof of Proposition 5.7.

The proof of Theorem 5.8 is the direct application of Theorem 5.1 with the choice of $p = 1$, setting $\delta' = \delta/(2K)$, and considering the probability of the event $\mathcal{E}_K$.

□

5.4 Discussion of the Main Result

Due to the dynamical nature of the MDP and AVI (and RFQI), the upper bound of Theorem 5.8 is more complicated than similar bounds in supervised learning. For instance, the effect of sample distribution ν on the quality of the final policy measured w.r.t. ρ is entangled with the dynamics of MDP itself (Section 5.3.1). Also the iterative nature of AVI relates the regression problem of the fitting procedure at any iteration to the solutions of earlier iterations. This effect shows itself in both the function approximation error (Section 5.3.3) and the smoothness of the target function (Section 5.3.4). These effects were obscure in the work of Munos and Szepesvári [2008] because of their more conservative analysis approach.

To better understand the behavior of RFQI, we explain the main terms of the upper bound in Theorem 5.8.

5.4.1 Error of the Fitting Procedure

The bounds on $\|Q_{k+1} - T^* Q_k\|_\nu^2$ in Propositions 5.6 and 5.7 have three terms. The term with $O(m_k^{-\frac{1}{1+\alpha_k}})$ behavior quantifies the estimation error while the terms $\|Q' - T^{*(k+1)} Q_0\|_\nu^2$ and $\sum_{i=0}^{k-1} \gamma^i b_{k-1-i}^2$ bound the function approximation error.

Estimation Error

The estimation error has the upper bound of (cf. (5.11) when Assumption A12 is used)

$$B\, c'_k(Q_{\max}) \left[L_R + \gamma L_P \left(\frac{L'_R(\delta')}{1 - L'_P(\delta')} + \left(J_0(Q_0) - \frac{L'_R(\delta')}{1 - L'_P(\delta')} \right) [L'_P(\delta')]^k \right) \right]^{\frac{2\alpha_k}{1+\alpha_k}} \times$$
$$m_k^{-\frac{1}{1+\alpha_k}} \left[\frac{\log(m_k \vee \frac{K}{\delta})}{\bar{\beta}_1} \right]^3 .$$

This upper bound shows the effect of the capacity of the function space $\mathcal{F}_k^{|\mathcal{A}|}$ and the smoothness of the target function.

Capacity of the function space. The effect of the number of samples and the capacity of the function space on the estimation error is $O(m_k^{-\frac{1}{1+\alpha_k}} \log^3(m_k))$. Disregarding the logarithmic term this is known to be the minimax optimal rate for i.i.d. inputs under the assumption that $\mathcal{F}_k$ has a packing entropy in the same form as in the upper bound of Assumption A7 [Yang and Barron, 1999]. This indicates that the effect of dependency in the input process is asymptotically negligible for exponential mixing processes. Since by setting $\gamma = 0$, the value-estimation task of an RL/Planning problem reduces to a regression problem, RL/Planning problems are superset of regression problems, and as a result this error bound is also optimal for the value-estimation task of Planning/RL problems.

Smoothness of the target function. The estimation error also depends on the smoothness of the target function. As discussed in Section 5.3.4, a deterministic upper bound on the smoothness of the target function can be obtained based on some intrinsic properties of the MDP, characterized by L_P and L_R (Assumption A11), and the parameter c defined in Proposition 5.5.

If for a constant $\delta' = \delta/(2K)$ the value of $L'_P(\delta') = c\gamma L_P \sqrt{\ln(1/\delta')}$ is larger than 1 and $\gamma > 0$, the smoothness term is dominated by $O([L'_P(\delta')]^k)$ for large values of k, which means that the target function of the later iterations can potentially become exponentially

non-smooth. On the other hand if L'_P is smaller than 1, the smoothness term is $O(L_P[L_R + \gamma L_P J_0(Q_0)])$ (see (5.11)). For a fixed confidence parameter δ, this smoothness is a finite constant and shows that the smoothness of the target function is upper bounded during all iterations.

It is noticeable that when $\gamma = 0$, the smoothness term behaves like $O(L_R)$, which is the smoothness of the reward function (see Assumption A11). This is expected as for $\gamma = 0$, we are essentially learning the reward function.

Function Approximation Error

The first term in the function approximation error is $\inf_{Q' \in \mathcal{F}_k^{|\mathcal{A}|}} \|Q' - (T^*)^{(k+1)} Q_0\|_\nu^2$, which is similar to the function approximation error in regression problems with the difference that the target is changing at each iteration. If the function space $\mathcal{F}_k^{|\mathcal{A}|}$ is not rich enough to approximate $T^{*(k+1)} Q_0$, this term suggests that the performance might be poor.

The second term of the function approximation error has $O(\sum_{i=0}^{k-1} \gamma^i b_{k-1-i}^2)$ behavior and shows the dependency of the function approximation error on the weighted sum of errors at the previous iterations. It indicates that a large error at previous iterations would cause a function approximation error in later iterations, though because of the discounting the effect shall become negligible. These two effects are obscure when one uses the inherent Bellman error (5.8) to bound the function approximation error [Munos and Szepesvári, 2008].

5.4.2 Influence of the Fitting Errors on the Resulting Policy

The main terms of the upper bound (5.13) are $\mathcal{E}(b_0, \dots, b_{K-1}; r)$ and $C_{\mathrm{VI},\rho,\nu}(K; r)$. Recalling that each b_k is an upper bound on $\|\varepsilon_k\|_\nu$, the term $\mathcal{E}(b_0, \dots, b_{K-1}; r)$ indicates how the fitting errors $(\varepsilon_k)_{k=0}^{K-1}$ influence the quality of the resulting policy. The term $C_{\mathrm{VI},\rho,\nu}(K; r)$ describes how the intrinsic properties of the MDP influence the error propagation. We have already discussed these two terms in Section 3.4 and compared it with previous work such as Munos [2007], so we avoid repeating the whole discussion here. Briefly speaking, this bound indicates that the errors of later iterations are more influential to the performance loss of the resulting policy and the effect of the intrinsic properties of the MDP and distributions ρ and ν is through the expectation of the squared Radon-Nikodym of the future state-action distributions w.r.t. the performance measuring distribution ρ.

5.5 Sparsity Regularities and l_1-Regularization

The RFQI algorithm introduced in Section 5.2 is indeed more general than the particular RKHS-based formulation of (5.2); it can be used with other choices of function space $\mathcal{F}^{|\mathcal{A}|}$ and regularizers J. In this section, we briefly describe one such possibility.

A promising class of candidate function spaces is the class of functions defined by wavelets [Antoniadis, 2007] or other over-complete dictionaries. Wavelets and over-complete dictionaries are intriguing because when chosen properly, they capture spatial irregularity and heterogeneity, such as spikes, that may occur in the action-value function. Wavelets are closely related to the Besov spaces, which are a large family of function spaces with a more general notion of smoothness compared to the smoothness used in the definition of Hölder or Sobolev spaces [Donoho and Johnstone, 1998].

To have smoothness-adaptive estimators for wavelets in the scale of Besov spaces, one may use either shrinkage-based estimators such as SureShrink [Donoho and Johnstone, 1995] or a regularization-based estimator such as the l_1-regularization-based one. These two types of estimators behave similarly for wavelets with orthogonal basis [Antoniadis, 2007]. In the rest of this section, we only briefly discuss the l_1-regularization-based formulation.

To design estimators based on wavelets or over-complete dictionaries, one possibility is to use the l_1-regularized least-squares regression, which is known as LASSO (Least Absolute

Shrinkage and Selection Operator) in the statistics and machine learning literature [Tibshirani, 1996]. Let us define the function space $\mathcal{F}^{|\mathcal{A}|}(p)$ as

$$\mathcal{F}^{|\mathcal{A}|}(p) = \{\Phi^{[p]}(\cdot,\cdot)^\top \theta | \theta \in \mathbb{R}^p, \Phi^{[p]}(\cdot,\cdot) : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^p\}$$

where $\Phi^{[p]}(\cdot,\cdot)$ is the feature vector defined by the first p wavelets or over-complete basis functions in some predetermined ordering. A natural ordering of basis functions for wavelets is to start from low-resolution "mother" and "scale" basis functions and move toward higher-resolution terms. For $Q(\cdot,\cdot;\theta) \in \mathcal{F}^{|\mathcal{A}|}(p)$, the regularizer would be $J(Q(\cdot,\cdot;\theta)) = |\theta|^\top \mu = \sum_{i=1}^p \mu_i |\theta_i|$, where $\mu \geq 0$ is the regularizer's weight vector. The vector μ imposes the prior belief about the contribution of each component of the basis functions $\Phi^{[p]}$. For wavelets choosing regularizer weights μ to decay exponentially according to the resolution-level of the wavelet coefficients is a natural choice in a Besov space, see e.g., Qu et al. [2009].

For this particular choice of function space $\mathcal{F}^{|\mathcal{A}|}$ and regularizer J, the optimization problem (5.1) would be

$$\theta_{k+1} = \operatorname*{argmin}_{\theta \in \mathbb{R}^{p_k}} \left\| Q(X_i, A_i; \theta) - \hat{T}^* Q_k(X_i, A_i; \theta_k) \right\|^2_{\mathcal{D}^{(k)}} + \lambda_k |\theta|^\top \mu,$$
$$Q_{k+1}(\cdot,\cdot;\theta_{k+1}) = \Phi^{[p_k]}(\cdot,\cdot)^\top \theta_{k+1}.$$

One should note that in addition to LASSO, there are other approaches to estimate a sparse function for a regression problem. Examples are SureShrink [Donoho and Johnstone, 1995], *adaptive LASSO* [Zou, 2006], *Adaptive Forward-Backward Greedy Algorithm* [Zhang, 2009b], and *Elastic net* [Zou and Hastie, 2005].

To provide an error upper bound for the l_1-regularized RFQI, similar to Theorem 5.8, two parts of the current analysis in Section 5.3 should be revised. The first is to provide a modified version of Theorem 5.2 (Section 5.3.2) for LASSO or any other l_1-regularized regression estimator. The second is to have a result similar to Proposition 5.4 that relates the l_1-norm of the estimate (i.e., $|\theta_{k+1}|^\top \mu$) to the l_1-norm of weights describing the function $\Pi_{\mathcal{F}_k} T^* Q_k$ in the function space $\mathcal{F}_k^{|\mathcal{A}|}$. In other words, we require a result that relates the smoothness of the estimate to the smoothness of the target function of the RFQI procedure.

One can indeed derive such results. A possible approach is to use the covering number result of Zhang [2002, Theorem 3]. This result, with the difference of a logarithmic factor of $\log(2p+1)$, satisfies Assumption A7 with the choice of $\alpha = 1$. Therefore, both Theorems 5.2 and Proposition 5.4 should hold without much modification. Nevertheless, we conjecture that this approach is not completely satisfactory as the squared error convergence rate would be slow and in the order of $O(m_k^{-1/2})$ in the setup of Theorem 5.2. The reason is that the covering number result of Zhang [2002] is quite generic and does not exploit any geometrical property of the function space. On the other hand, if we have some extra assumption about the basis functions $\Phi^{[p]}$, such as some form of the Restricted Isometric Property, we may get faster than $O(m_k^{-1/2})$ rates, see e.g., Zhang [2009a].

5.6 Conclusion and Future Work

In this work we proposed to use regularized regression, a powerful technique in the nonparametric supervised learning literature, in the AVI procedure in order to solve RL/Planning problems with large state spaces. Our formulation of RFQI was general and could incorporate various function spaces and regularization functionals (Section 5.2). This includes a broad class of RKHS, over-complete dictionaries and wavelets, neural networks, and of course parametric models. We specifically focused on the RKHS formulation as it has advantages such as the generality to work with different input domains and the ease of choosing/changing the kernel function and consequently the function space.

A considerable part of this work has been devoted to analyzing the statistical behavior of RFQI (Section 5.3). We provided an error upper bound on the performance loss of the resulting policy compared with the optimal policy's (Theorem 5.8). The error bound indicates

the role of the sample size, complexity of function space to which the estimate belongs (quantified by its metric entropy), function approximation error, and the intrinsic properties of the MDP such as the behavior of concentrability coefficients and the smoothness-expansion property of the Bellman optimality operator. We discussed the interpretation of our result in Section 5.4, so to avoid duplication we do not dwell on it here anymore.

This work opens up several possibilities for future research. We mention some of them here.

Applications. RFQI is a flexible algorithm that may be applied to many real-world RL/Planning problems. Nevertheless, there has not been many real-world applications of it yet. The only exception is the work of Farahmand et al. [2009c] who apply RFQI to the visual-servoing task for the robotic arm manipulation (see also Farahmand et al. [2008] for some experiments on the effect of regularization coefficient and the choice of kernel on the performance in a toy problem). More real-world applications of RFQI is the topic of future work.

Computational Considerations. RFQI method is simple to implement as it is essentially a repeated application of a regression algorithm. For large datasets, however, extra care is required. A naive implementation of, say, an RKHS-based regularized regression requires inversion of matrices with the size equal to the number of samples. This requires the computation time of $O(\frac{n^3}{K^2})$, which is prohibitive for large sample sizes. This kind of computational problem, however, is common to many nonparametric methods. We mention three approaches to design more efficient algorithms.

One possible approach to reduce computational cost is to use *sparsification* techniques developed in the kernel-based learning literature and have been used in the RL/Planning literature [Engel et al., 2005; Jung and Polani, 2006; Xu et al., 2007]. The idea of these methods is to retain a small subset of "representative" data samples as the active kernel bases. As a result, the size of matrices involved in the computation would be reduced. Refer to Section 8.3 of Rasmussen and Williams [2006] for more information.

An efficient way to solve large-scale linear systems is to use iterative methods such as conjugate gradient algorithm. The bottleneck of these types of algorithms is the matrix-vector multiplication that costs $O(n'^2)$ with n' being the size of the vector, e.g., number of sample points in each iteration. One elegant approach to reduce the complexity of matrix-vector multiplication is to use Fast Multipole Methods (FMM) [Beatson and Greengard, 1997] and its variants such as Fast Gauss Transform to reduce the computation cost to $O(n' \log n')$ or better at the cost of some small, but controlled, error [Yang et al., 2004]. These methods are particularly efficient for low-dimensional problems.

The other possibility is to use stochastic gradient methods to solve the corresponding optimization problem. This is especially appealing in the light of results such as Bottou and Bousquet [2008], which show that given a fixed amount of computation time, the generalization error resulting from learning with stochastic gradient methods as the optimizer might be less than that of gradient descent methods.

Continuous Action Space. Another important question, especially for practical real-world applications, is how to extend RFQI to deal with continuous action MDPs. One difficulty of extending our current result to the continuous-action one is finding the maximizing action at each state, which is needed to estimate the Bellman optimality operator. Except in special cases, this cannot be done exactly. Instead one may use a local search, similar to what is done by Xu et al. [2010]. To analyze this inexact policy improvement some parts of the theory, especially the error propagation result (Section 5.3.1), should be modified. Moreover, it also seems that 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].

Model Selection. The successful application of any RL/Planning algorithm, including RFQI, depends on the proper choice of its parameters. For the case of RFQI, we are faced with the choice of $(\mathcal{F}_k^{|\mathcal{A}|})$ and the corresponding regularization parameters (λ_{m_k}) (5.1). The optimal choice of these parameters, however, are problem-dependent and unknown.

To see the issue more clearly, focus on the optimal choice of regularization coefficient, which according to Theorem 5.8 should be $\lambda_{m_k} = B\,[\frac{1}{m_k J(\Pi_{\nu,\mathcal{F}_k^{|\mathcal{A}|}} T^* Q_k)}]^{1/(1+\alpha_k)}$. This choice depends on unknown parameters such as $J(\Pi_{\nu,\mathcal{F}_k^{|\mathcal{A}|}} T^* Q_k)$ and the smoothness order of $\mathcal{F}_k^{|\mathcal{A}|}$ described by α_k. In general, these values are not known. The right approach to solve this issue is to choose the parameters of the algorithm data-dependently by a model selection procedure. We introduce such an algorithm in Chapter 7. Here, we only focus on an alternative approach that benefits from the specifics of the RFQI procedure.

The problem of model selection seems to be easier for AVI procedures, such as RFQI, compared to when we are only given a bunch of action-value functions and have to choose their best. The reason is that AVI solves a sequence of regression problems, so one may simply perform a model selection procedure at each iteration of AVI. To be more concrete, consider the k^{th} iteration of RFQI. Construct p_1 function spaces $\mathcal{F}_k^{|\mathcal{A}|^{(i)}}$ with different smoothness orders $\alpha^{(i)}$ $(i = 1, \dots, p_1)$. Also construct p_2 values of $\lambda_k^{(j)}$ $(j = 1, \dots, p_2)$. The result is $P = p_1 \times p_2$ potential models. Now estimate $Q_{k+1}^{(l)}$ for $l = 1, \dots, P$, and select the best combination by the aid of any model selection approach for regression problems such as the cross-validation procedure [Arlot and Celisse, 2009] or the complexity regularization-based approach [Wegkamp, 2003]. This leads to an estimate whose convergence bound has the optimal order and scales with the actual roughness $J(\Pi_{\mathcal{F}_k} T^* Q_k)$.

Computational issues aside, one should perform the model selection process at each iteration of RFQI. This is important because the appropriate regularization coefficients and the function spaces may change during iterations.

There are, however, some subtleties. First issue is that because the data samples distribution ν is different from the performance measuring distribution ρ, finding the best model according to ν is not necessarily the best choice. Second issue is that of computational cost. If we are computationally limited, we may not want to perform model selection at all K iterations of RFQI. As discussed in Section 3.4, the errors at later iterations are more important than the errors at earlier ones. How to optimally distribute the points of performing model selection during K iterations is a practically important question.

l_1-Regularization. In Section 5.5, we briefly mentioned how one may analyze l_1-based RFQI. Fully developing this theory seems to require extending the current LASSO error bounds to mixing processes. This is a topic for future research.

Influence of the MDP on the Smoothness and Function Approximation Error. An open theoretical question is to characterize the properties of MDP that determine the values of L_P and L_R in Assumption A11. In Proposition 6.16 (Appendix 6.G), we prove the conditions that for the certain class of MDPs, which we call *convolutional* MDPs, Assumption A11 holds. Briefly speaking, the conditions are 1) the transition probability kernel should have a finite gain (in control theoretic sense) in its frequency response, and 2) the reward function should be smooth. Another question is to determine the influence of the Bellman optimality operator and the function space on the function approximation error $\inf_{Q' \in \mathcal{F}_k^{|\mathcal{A}|}} \|Q' - (T^*)^{(k+1)} Q_0\|_\nu$ in Theorem 5.3 and consequently Theorem 5.8.

Other Technical Questions. Some technical questions have not yet been addressed. One open question is how to extend Proposition 5.4 to mixing processes. We conjecture that it should also hold for mixing processes too but a rigorous proof is needed. Another technical issue is regarding the oversmoothing assumption stated as Assumption A12 and the conditions under which it holds.

Appendices

5.A Error Bounds for Regularized Regression: Proofs for Section 5.3.2

To extend the result of Chapter 4 to the multivariate regression setting of RFQI, we need an upper bound on the metric entropy of $\mathcal{F}_k^{|\mathcal{A}|}$. Propositions 5.9 and 5.10 allow us to relate the metric entropy of $\mathcal{F}_k^{|\mathcal{A}|}$ to that of $\mathcal{F}_k$. These results indicate that the effect of having finitely many actions $\mathcal{A}$ is only in the multiplicative constant of the metric entropy. Theorem 5.2 is the direct consequence of Theorem 4.5 of Chapter 4 and these two propositions.

Proposition 5.9. *Consider a function space $\mathcal{F}_+ \subset \mathcal{F} : \mathcal{X} \to \mathbb{R}$, a finite set $\mathcal{A}$, and the corresponding function space $\mathcal{F}_+^{|\mathcal{A}|} = \underbrace{\mathcal{F}_+ \times \cdots \times \mathcal{F}_+}_{|\mathcal{A}| \text{ times}}$. Then for any $((x_1, a_1), \ldots, (x_n, a_n)) \in (\mathcal{X} \times \mathcal{A})^n$, we have*

$$\log \mathcal{N}_2(u, \mathcal{F}_+^{|\mathcal{A}|}, (x,a)_{1:n}) \leq |\mathcal{A}| \log \mathcal{N}_2(\frac{u}{\sqrt{|\mathcal{A}|}}, \mathcal{F}_+, x_{1:n}).$$

Proof of Proposition 5.9. For any $V \in \mathcal{F}$, denote the empirical norm by $\|V\|_{x_{1:n}}^2 = \frac{1}{n}\sum_{i=1}^n |V(x_i)|^2$, and for any $Q \in \mathcal{F}^{|\mathcal{A}|}$, denote $\|Q\|_{(x,a)_{1:n}}^2 = \frac{1}{n}\sum_{i=1}^n |Q(x_i, a_i)|^2$. Now suppose that for any $a \in \mathcal{A}$, the set $\{Q_1(\cdot, a), \ldots, Q_{N_a}(\cdot, a)\}$ is an δ-covering of $\mathcal{F}_+$ w.r.t. the empirical norm $\|Q(\cdot, a)\|_{x_{1:n}}^2$. Therefore for any $Q(\cdot, a) \in \mathcal{F}_+$, there exists a member of the aforementioned set that has a distance less than δ to any $Q(\cdot, a) \in \mathcal{F}_+$. Now pick $Q_1, Q_2 \in \mathcal{F}_+^{|\mathcal{A}|}$. We have $\|Q_1 - Q_2\|_{(x,a)_{1:n}}^2 = \frac{1}{n}\sum_{i=1}^n |Q_1(x_i, a_i) - Q_2(x_i, a_i)|^2 \leq \sum_{a\in\mathcal{A}} \|Q_1(\cdot, a) - Q_2(\cdot, a)\|_{x_{1:n}}^2$. Therefore, if $\max_{a\in\mathcal{A}} \|Q_1(\cdot, a) - Q_2(\cdot, a)\|_{x_{1:n}}^2 \leq \delta^2$, then $\|Q_1 - Q_2\|_{(x,a)_{1:n}}^2 \leq |\mathcal{A}|\delta^2$, and the set

$$\prod_{a\in\mathcal{A}} \Big\{Q_1(\cdot, a), \ldots, Q_{N_a}(\cdot, a)\Big\}$$

makes an $\sqrt{|\mathcal{A}|}\delta$-covering of $\mathcal{F}_+^{|\mathcal{A}|}$. The cardinality of this covering is $N_1 \times \ldots \times N_{|\mathcal{A}|} = [\mathcal{N}_2(\delta, \mathcal{F}_+, x_{1:n})]^{|\mathcal{A}|}$. Set $\delta = \frac{u}{\sqrt{|\mathcal{A}|}}$ to get the result. □

The following proposition is the direct consequence of Proposition 5.9 applied to the ball defined in Assumption A7. In this proposition, we let $\mathcal{B}_{k,R}^{|\mathcal{A}|} \triangleq \{Q \in \mathcal{F}_k^{|\mathcal{A}|} : J_k^2(Q) \leq R^2\}$.

Proposition 5.10. *Let Assumptions A6 and A7 hold. There exists a constant $C' > 0$ such that for all $(x, a)_{1:n} \in \mathcal{X} \times \mathcal{A}$, we have*

$$\log \mathcal{N}_2(u, \mathcal{B}_{k,R}^{|\mathcal{A}|}, (x,a)_{1:n}) \leq C' \left(\frac{R}{u}\right)^{2\alpha_k}.$$

In particular, one can choose $C' = |\mathcal{A}|^{1+\alpha_k} C$.

Proof. By Assumption A6, $\mathcal{B}_{k,R}^{|\mathcal{A}|} \subset \prod_{a\in\mathcal{A}} \mathcal{B}_{k,a,R}$. The result is a direct consequence of Assumption A7 and Proposition 5.9. □

5.B The Behavior of the Function Approximation Error: Proofs for Section 5.3.3

In this section, we prove Theorem 5.3. We first present a lemma which shows that the Bellman optimality operator is Lipschitz when viewed as an operator of the Banach space of action-value functions equipped with $\|\cdot\|_\nu$.

Lemma 5.11. *For any given $Q_1, Q_2 \in \mathcal{F}^{|\mathcal{A}|}$, we have $\|T^*Q_1 - T^*Q_2\|_\nu \le \gamma\, C_{AE}(\nu; P)\, \|Q_1 - Q_2\|_\nu$.*

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

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

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

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

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

Proof of Theorem 5.3. Let $Q_0, \ldots, Q_{K-1}$ be action-value functions, $\varepsilon_k = T^*Q_k - Q_{k+1}$, $b_k = \|\varepsilon_k\|_\nu$. Our goal is to bound $\inf_{Q'\in\mathcal{F}_k^{|\mathcal{A}|}} \|Q' - T^*Q_k\|_\nu$. For this, pick any $Q' \in \mathcal{F}_k^{|\mathcal{A}|}$. Then, by the triangle inequality,

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

therefore, it remains to upper bound $\|(T^*)^{k+1}Q_0 - T^*Q_k\|_\nu$. Since by Lemma 5.11, T^* is $L \triangleq \gamma C_{\mathrm{AE}}(\nu; P)$-Lipschitz w.r.t. $\|\cdot\|_\nu$, we have $\|(T^*)^{k+1}Q_0 - T^*Q_k\|_\nu \le L\|(T^*)^kQ_0 - Q_k\|_\nu$. Using the definition of ε_k, $\|(T^*)^kQ_0 - Q_k\|_\nu = \|(T^*)^kQ_0 - (T^*Q_{k-1} - \varepsilon_{k-1})\|_\nu \le \|(T^*)^kQ_0 - T^*Q_{k-1}\| + \|\varepsilon_{k-1}\|_\nu \le L\|(T^*)^{k-1}Q_0 - Q_{k-1}\| + \|\varepsilon_{k-1}\|_\nu$. Finishing the recursion gives

$$\|(T^*)^kQ_0 - Q_k\|_\nu \le \|\varepsilon_{k-1}\|_\nu + L\|\varepsilon_{k-2}\|_\nu + \ldots + L^{k-1}\|\varepsilon_0\|_\nu.$$

Combining the inequalities obtained so far, we get

$$\|Q' - T^*Q_k\|_\nu \le \|Q' - (T^*)^kQ_0\|_\nu + \sum_{i=0}^{k-1} L^{k-i}\|\varepsilon_i\|_\nu,$$

from which the desired statement follows immediately. □

5.C The Behavior of the Smoothness: Proofs for Section 5.3.4

Proof of Proposition 5.5. According to our assumptions, for $k = 0, 1, \ldots, K-1$ we have

$$\begin{aligned}
J_k(Q_{k+1}) &\le cJ_k(\Pi_{\nu,\mathcal{F}_k^{|\mathcal{A}|}} T^*Q_k)\sqrt{\ln(1/\delta)} \\
&\le (cL_R\sqrt{\ln(1/\delta)}) + (c\gamma L_P\sqrt{\ln(1/\delta)})J_{k-1}(Q_k) \\
&= L'_R(\delta) + L'_P(\delta)J_{k-1}(Q_k)
\end{aligned}$$

where we used Assumption A11 in the second inequality, and the definitions of L'_R and L'_P in the last step. Now, consider the recursion

$$\bar{J}_{k+1} = L'_R(\delta) + L'_P(\delta)\bar{J}_k,$$

where $\bar{J}_0 = J_{-1}(Q_0) = J_0(Q_0)$. By induction, we see that $J_{k-1}(Q_k) \leq \bar{J}_k$ holds for $0 \leq k \leq K-1$. By solving the recursion, we get

$$\bar{J}_k = [L'_P(\delta)]^k \bar{J}_0 \, \mathbb{I}(k \geq 0) + \frac{L'_R(\delta)}{1 - L'_P(\delta)} \left\{1 - [L'_P(\delta)]^k\right\} \mathbb{I}(k \geq 1).$$

Another application of Assumption A11 leads to the desired result. □

Chapter 6

Regularized Policy Iteration Algorithm

6.1 Introduction

In this chapter, we provide two regularization-based nonparametric API algorithms, namely *Regularized Least-Squares Policy Improvement* **(REG-LSPI)** and *Regularized Bellman Residual Minimization* **(REG-BRM)** to solve RL/Planning problems with large state spaces. These are flexible methods that upon the proper choice of their parameters can efficiently deal with RL/Planning problems with large state spaces. Both REG-BRM and REG-LSPI are formulated as coupled optimization problems (Section 6.3) for which we provide closed-form solutions when the estimated action-value function is chosen from the family of RKHSs (Section 6.3.1).[1][2]

The theoretical contribution of this work (Section 6.4) is to analyze the statistical properties of REG-LSPI and to provide upper bounds on the policy evaluation error and as a consequence the quality of the resulting policy and its relation to the performance of the optimal policy (Theorem 6.6). 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. We see that the dependence of the policy evaluation error bound on the number of samples is minimax optimal.

We overview API in some detail in Section 6.2 and then focus on the regularized API algorithms in Section 6.3. We analyze the statistical behavior of REG-LSPI in Section 6.4.

6.2 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

[1] This chapter is the result of the collaboration of the author with Csaba Szepesvári, Mohammad Ghavamzadeh, and Shie Mannor.

[2] This chapter is slightly revised from the original PhD Dissertation submitted in September 2011. The changes will be noted by footnotes.

repeating this process again and generating a sequence of policies and their corresponding approximate action-value functions $Q_0 \to \pi_1 \to Q_1 \to \pi_2 \to \cdots$.[3]

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,[4] 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 that the distribution of (X_i, A_i) in the data samples is typically different from the "ideal" distribution, i.e., a distribution that would be used when the learned policy is evaluated. This causes a problem since the methods must be able to handle this mismatch of distributions (a number of works in the supervised learning literature consider this scenario, see e.g., Ben-David et al. [2006]; Mansour et al. [2009]). 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 6.3.

6.2.1 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 another 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*. If this norm $\|Q - T^*Q\|$ is small, then Q is a good approximation of Q^*. An intuitive consequence is that the value function of the greedy policy w.r.t. Q, that is $V^{\hat{\pi}(\cdot;Q)}$, should also in some sense be close to the optimal value function V^*. This intuition is indeed correct and has been formalized when the Bellman error is defined by either the L_∞-norm [Williams and Baird, 1994] or an L_p-norm (Munos [2003]; Antos et al. [2008b]; Farahmand et al. [2010] and Theorem 6.5 of this work). For instance, an early result such as Williams and Baird [1994] states that

$$\left\| V^* - V^{\hat{\pi}(\cdot;Q)} \right\|_\infty \leq \frac{2}{1-\gamma} \left\| Q - T^*Q \right\|_\infty .$$

The supremum norm, however, is too conservative in many practical situations. This is especially the case when we are dealing with large state spaces for which one must use function approximation. Point-wise convergence results in supervised learning theory usually requires stronger conditions on the sampling distribution, and we doubt if it is a good idea to use them in the RL/Planning context either.

To make this point clearer, consider a situation where the agent uses Q as an approximation to the optimal action-value function Q^*, and uses $\hat{\pi}(\cdot;Q)$ as its policy. Moreover, we measure the performance of the agent w.r.t. the initial-state evaluation distribution ρ. That is for a given π, we measure

$$V(\pi;\rho) = \int_{\mathcal{X}\times\mathcal{A}} Q^\pi(x,a) d\rho(x,a).$$

What we are interested in is the performance loss (regret) of the policy π compared to the optimal one, i.e., $V(\pi^*;\rho) - V(\hat{\pi}(\cdot;Q);\rho)$. For example, if ρ is the Lebesgue measure on

[3] 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.

[4] 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, $\mathcal{L} = \mathbf{I} - P^\pi$, $g(\cdot) = r(\cdot, \pi(\cdot))$, and $f = Q^\pi$.

$\mathcal{X} \times \mathcal{A}$ (uniform distribution for compact $\mathcal{X} \times \mathcal{A}$), it indicates that what is important to the designer is that the agent performs equally well for all initial state-action. Now consider that $Q = Q^\pi$ in all state space except a ρ-tiny region $\mathcal{X}_1 \times \mathcal{A}_1 \subset \mathcal{X} \times \mathcal{A}$, i.e., $\rho(\mathcal{X}_1 \times \mathcal{A}_1) \ll 1$. In $\mathcal{X}_1 \times \mathcal{A}_1$, Q is largely different from Q^*. Here, $\|Q - T^*Q\|_\infty$ has a large value, however, the performance of the agent following $\hat{\pi}(\cdot; Q)$ is in expectation w.r.t. ρ very close to the optimal performance.

A more natural choice is to use a weighted L_p-norms such as the L_2-norm to measure the magnitude of the Bellman residual. This leads to a tractable optimization problem and enables a connection to regression function estimation [Györfi et al., 2002]. More importantly, results such as Munos [2003]; Antos et al. [2008b]; Farahmand et al. [2010] and Theorem 6.5 of this work show that minimizing the L_p-norm of the Bellman residual $\|Q - T^*Q\|_{p,\nu}$ (with ν being the sampling distribution and $p \geq 1$) leads to minimizing an upper bound on the performance loss $\left\|Q^* - Q^{\hat{\pi}(\cdot;Q)}\right\|_{p',\rho}$ (for some distribution ρ and a well-specified $p' \geq 1$). In the special case of $p = 1$, $\left\|Q^* - Q^{\hat{\pi}(\cdot;Q)}\right\|_{1,\rho}$ has an appealing interpretation: It is the expected regret of following policy $\hat{\pi}(\cdot; Q)$ instead of the optimal policy when the initial state-action distribution is ρ. More on this issue in Section 6.4.2.

The BRM algorithm is defined as the procedure minimizing the following loss function:

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

where ν is the distribution of state-actions in the input data. Using the empirical L_2-norm defined in (2.5) with samples $\mathcal{D}_n$ defined in (2.6), and by replacing $(T^\pi Q)(X_t, A_t)$ with the empirical Bellman operator (Definition 2.8), the empirical estimate of $L_{BRM}(Q; \pi)$ can be written as

$$\begin{aligned} \hat{L}_{BRM}(Q; \pi, n) &\triangleq \left\|Q - \hat{T}^\pi Q\right\|_{\mathcal{D}_n}^2 \\ &= \frac{1}{n} \sum_{t=1}^{n} \left[Q(X_t, A_t) - \Big(R_t + \gamma Q\big(X'_t, \pi(X'_t)\big)\Big)\right]^2 . \end{aligned} \tag{6.1}$$

Nevertheless, it is well-known that $\hat{L}_{BRM}$ is *not* an unbiased estimate of L_{BRM} when the MDP is not deterministic [Sutton and Barto, 1998; Lagoudakis and Parr, 2003; Antos et al., 2008b]: For any fixed Q,

$$\begin{aligned} \mathbb{E}\left[\hat{L}_{BRM}(Q; \pi, n)\right] &= \mathbb{E}\left[\left\|Q - \hat{T}^\pi Q\right\|_{\mathcal{D}_n}^2\right] \\ &= \|Q - T^\pi Q\|_\nu^2 + \mathbb{E}\left[\left\|T^\pi Q - \hat{T}^\pi Q\right\|_{\mathcal{D}_n}^2\right] \\ &\neq L_{BRM}(Q; \pi). \end{aligned} \tag{6.2}$$

The reason, as is evident in (6.2), is that stochastic transitions/rewards lead to a non-vanishing variance term because $T^\pi Q \neq \hat{T}^\pi Q$. This extra term can be problematic. Whenever the dynamical system has stochastic transitions, this variance term is not fixed and is Q-dependent. Therefore, the minimizing of $\hat{L}_{BRM}$ is not the same as the minimizer of L_{BRM} – even in the ideal situation of not having any estimation error.

One suggestion to deal with this problem is to use double-sampling to estimate $\hat{L}_{BRM}$. According to this proposal, from each state-action pair in the sample, we require to have at least two independent next-state samples, see e.g., Sutton and Barto [1998]. Nevertheless, this suggestion may not be practical in many cases. The luxury of having two next-state samples is not available in the RL scenario. Even if we have a generative model of the environment, as we do in the planning scenario, the result would not be sample-efficient, which is important when generating new samples is costly. When the generative model is available and the double-sampling is not an issue, Maillard et al. [2010] analyze BRM for the finite linear function spaces.

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

$$L_{BRM}(Q, h; \pi) = L_{BRM}(Q; \pi) - \|h - T^\pi Q\|_\nu^2 , \tag{6.3}$$

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

$$Q_{BRM} = \operatorname*{argmin}_{Q \in \mathcal{F}^{|\mathcal{A}|}} \sup_{h \in \mathcal{F}^{|\mathcal{A}|}} L_{BRM}(Q, h; \pi), \tag{6.4}$$

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 (6.4) is equivalent to the set of following coupled (nested) optimization problems:

$$\begin{aligned} h(\cdot; Q) &= \operatorname*{argmin}_{h' \in \mathcal{F}^{|\mathcal{A}|}} \|h' - T^\pi Q\|_\nu^2 , \\ Q_{BRM} &= \operatorname*{argmin}_{Q \in \mathcal{F}^{|\mathcal{A}|}} \left[\|Q - T^\pi Q\|_\nu^2 - \|h(\cdot; Q) - T^\pi Q\|_\nu^2 \right] . \end{aligned} \tag{6.5}$$

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.,

$$\hat{h}_n(\cdot; Q) = \operatorname*{argmin}_{h \in \mathcal{F}^{|\mathcal{A}|}} \left\| h - \hat{T}^\pi Q \right\|_{\mathcal{D}_n}^2 , \tag{6.6}$$

$$\hat{Q}_{BRM} = \operatorname*{argmin}_{Q \in \mathcal{F}^{|\mathcal{A}|}} \left[\left\| Q - \hat{T}^\pi Q \right\|_{\mathcal{D}_n}^2 - \left\| \hat{h}_n(\cdot; Q) - \hat{T}^\pi Q \right\|_{\mathcal{D}_n}^2 \right] . \tag{6.7}$$

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

6.2.2 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 is 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 ν-weighted projection operator onto the space of admissible function $\mathcal{F}^{|\mathcal{A}|}$, i.e., $\Pi_\nu = \Pi_{\mathcal{F}_\nu^{|\mathcal{A}|}} : B(\mathcal{X} \times \mathcal{A}) \to B(\mathcal{X} \times \mathcal{A})$ is defined by $\Pi_{\mathcal{F}_\nu^{|\mathcal{A}|}} Q = \operatorname{argmin}_{h \in \mathcal{F}^{|\mathcal{A}|}} \|h - Q\|_\nu^2$ for $Q \in B(\mathcal{X} \times \mathcal{A})$. If the operator $\Pi_\nu T^\pi$ is a contraction operator, Banach fixed-point theorem implies that the combined operator has a unique fixed-point (Theorem B.3 in Appendix B.5).

Nevertheless, the operator $(\Pi_\nu T^\pi)$ is not contraction for arbitrary choice of ν (an exception is when ν is the stationary distribution induced by π). Therefore, when the distribution of samples $(X_t, A_t) \sim \nu$ is different from the stationary distribution induced by π, this equation does not necessarily have a unique fixed point.

To address this issue, one may define the LSTD solution as the minimizer of the L_2-norm between Q and $\Pi_\nu T^\pi Q$:

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

The minimizer of $L_{LSTD}(Q; \pi)$ is well-defined, and whenever ν is the stationary distribution of π, the solution to this optimization problem is the same as the solution to $Q = \Pi_\nu T^\pi Q$.

Algorithm 2 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 sample $\mathcal{D}_n^{(k)}$
  $\hat{Q}^{(k)} \leftarrow$ 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 $\pi_K(\cdot) = \hat{\pi}(\cdot; \hat{Q}^{(K-1)})$

The LSTD solution can therefore be written as the solution to the following set of coupled optimization problems:

$$h(\cdot; Q) = \operatorname*{argmin}_{h' \in \mathcal{F}^{|\mathcal{A}|}} \|h' - T^\pi Q\|_\nu^2 ,$$
$$Q_{LSTD} = \operatorname*{argmin}_{Q \in \mathcal{F}^{|\mathcal{A}|}} \|Q - h(\cdot; Q)\|_\nu^2 , \tag{6.8}$$

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 dataset $\mathcal{D}_n$ is

$$\hat{h}_n(\cdot; Q) = \operatorname*{argmin}_{h \in \mathcal{F}^{|\mathcal{A}|}} \left\| h - \hat{T}^\pi Q \right\|_{\mathcal{D}_n}^2 , \tag{6.9}$$

$$\hat{Q}_{LSTD} = \operatorname*{argmin}_{Q \in \mathcal{F}^{|\mathcal{A}|}} \left\| Q - \hat{h}_n(\cdot; Q) \right\|_{\mathcal{D}_n}^2 . \tag{6.10}$$

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 6.1a 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 (6.5) coincides with the LSTD solution (6.8) [Antos et al., 2008b]. The reason is that the first equation in both (6.5) and (6.8) finds the projection $\hat{h}_n(\cdot; Q)$ of $T^\pi Q$ to $\mathcal{F}^{|\mathcal{A}|}$, thus $\hat{h}_n(\cdot; Q) - T^\pi Q$ is perpendicular to $\mathcal{F}^{|\mathcal{A}|}$. Therefore, we can use Pythagorean theorem to get $\|Q - \hat{h}_n(\cdot; Q)\|^2 = \|Q - T^\pi Q\|^2 - \|\hat{h}_n(\cdot; Q) - T^\pi Q\|^2$. This implies that the second equations in (6.5) and (6.8) have the same solution.

6.3 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 6.2.1) and LSTD (Section 6.2.2) for the task of approximate policy evaluation.

The pseudo-code of the Regularized Policy Iteration algorithms is shown in Algorithm 2. The algorithm receives K (the number of API iterations), an initial action-value function

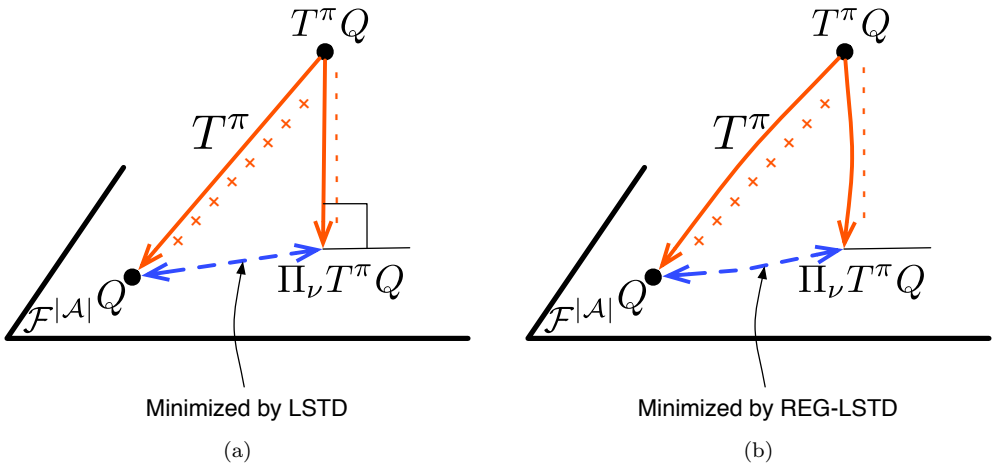

Figure 6.1: (a) This figure shows the loss functions minimized by the original BRM, the modified BRM, and the LSTD methods. The function space $\mathcal{F}^{|\mathcal{A}|}$ is represented by the plane. The Bellman operator T^π maps an action-value function $Q \in \mathcal{F}^{|\mathcal{A}|}$ to a function $T^\pi Q$. 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 $\mathcal{F}^{|\mathcal{A}|}$.

$\hat{Q}^{(-1)}$, 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., $\pi_k \leftarrow \hat{\pi}(\cdot; \hat{Q}^{(k-1)}) = \operatorname{argmax}_{a' \in \mathcal{A}} \hat{Q}^{(k-1)}(\cdot, a')$. 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} in each iteration in such a way that the agent ultimately achieves a better performance.[5] This 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., $\pi_{b_k} = \pi_b$. This leads to K independent datasets $\mathcal{D}_n^{(0)}, \ldots, \mathcal{D}_n^{(K-1)}$. 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 done by REG-LSTD/BRM, which will be discussed shortly. These procedures receive policy π_k, the training samples $\mathcal{D}_n^{(k)}$, the function

[5] There are various heuristics to choose the behavior policy π_{b_k}. One is to select a fixed stochastic stationary policy π_b in all iterations. Another is to use a policy based on the most recent estimate of the action-value function, i.e., $Q^{(k-1)}$. This can be the greedy policy w.r.t. $Q^{(k-1)}$ with some exploration. For instance, suppose $\Delta_\pi : \mathcal{X} \times \mathcal{A} \to [0, 1]$ is a function of state and action with the property that $\sum_{i=1}^{|\mathcal{A}|} \Delta_\pi(x, a_i) \le \varepsilon \le 1$ for all $x \in \mathcal{X}$. For a given deterministic policy $\pi : \mathcal{X} \to \mathcal{A}$, define the perturbed policy $\pi \oplus \Delta_\pi$ as the probability distribution of action selection at each state as

$$(\pi \oplus \Delta_\pi)(\cdot|x) \triangleq \begin{cases} a_i & \text{with probability } \Delta_\pi(x, a_i) \\ \pi(x) & \text{with probability } 1 - \sum_{i=1}^{|\mathcal{A}|} \Delta_\pi(x, a_i) \end{cases}$$

One may then define policy $\pi_{b_k} = \hat{\pi}(\cdot; Q^{(k-1)}) \oplus \Delta_\pi$ for some choice of Δ_π, e.g., $\Delta_\pi(\cdot, a) = \varepsilon/|\mathcal{A}|$ with $0 \le \varepsilon < 1$ (for all $a \in \mathcal{A}$).

space $\mathcal{F}^{|\mathcal{A}|}$, the regularizer J, and the regularization coefficients $(\lambda_{Q,n}^{(k)}, \lambda_{h,n}^{(k)})$, and return 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:

$$\hat{h}_n(\cdot; Q) = \operatorname*{argmin}_{h \in \mathcal{F}^{|\mathcal{A}|}} \left[\left\| h - \hat{T}^{\pi_k} Q \right\|_{\mathcal{D}_n}^2 + \lambda_{h,n}^{(k)} J^2(h) \right], \tag{6.11}$$

$$\hat{Q}^{(k)} = \operatorname*{argmin}_{Q \in \mathcal{F}^{|\mathcal{A}|}} \left[\left\| Q - \hat{T}^{\pi_k} Q \right\|_{\mathcal{D}_n}^2 - \left\| \hat{h}_n(\cdot; Q) - \hat{T}^{\pi_k} Q \right\|_{\mathcal{D}_n}^2 + \lambda_{Q,n}^{(k)} J^2(Q) \right], \tag{6.12}$$

where $J : \mathcal{F}^{|\mathcal{A}|} \to \mathbb{R}$ is the regularizer functional (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 (2.6). We call $J(Q)$ the smoothness of Q.

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

$$\hat{h}_n(\cdot; Q) = \operatorname*{argmin}_{h \in \mathcal{F}^{|\mathcal{A}|}} \left[\left\| h - \hat{T}^{\pi_k} Q \right\|_{\mathcal{D}_n}^2 + \lambda_{h,n}^{(k)} J^2(h) \right], \tag{6.13}$$

$$\hat{Q}^{(k)} = \operatorname*{argmin}_{Q \in \mathcal{F}^{|\mathcal{A}|}} \left[\left\| Q - \hat{h}_n(\cdot; Q) \right\|_{\mathcal{D}_n}^2 + \lambda_{Q,n}^{(k)} J^2(Q) \right]. \tag{6.14}$$

Note that the difference between (6.6)-(6.7) ((6.9)-(6.10)) and (6.11)-(6.12) ((6.13)-(6.14), respectively) is the addition of the regularizers $J^2(h)$ and $J^2(Q)$.

Unlike the non-regularized case described in Section 6.2, the solutions of REG-BRM and REG-LSTD are not the same. As a result of the *regularized* projection, (6.11) and (6.13), 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 6.1b).

One may ask why we have regularization terms in both optimization problems, as opposed to only in the projection term (6.13) (similar to the Lasso-TD algorithm Kolter and Ng 2009; Ghavamzadeh et al. 2011) or only in (6.14) (similar to Geist and Scherrer 2012; Ávila Pires and Szepesvári 2012). We discuss this question in Appendix 6.F. Briefly speaking, for large function spaces such as Sobolev spaces or RKHS with universal kernels, if we remove the regularization term in (6.13), the coupled optimization problems reduces to (unmodified) BRM, which is biased as discussed earlier; whereas if the regularization term in (6.14) is removed, the solution can be arbitrary bad due to overfitting.[6]

6.3.1 Closed-Form Solutions

In this section we provide a closed-form solution for (6.11)-(6.12) and (6.13)-(6.14) for two cases: 1) When $\mathcal{F}^{|\mathcal{A}|}$ is a finite dimensional linear space and $J(\cdot)$ 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., $J(\cdot) = \|\cdot\|_{\mathcal{H}}$.

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:

$$h(\cdot) = \boldsymbol{\phi}(\cdot)^\top \boldsymbol{u}, \quad Q(\cdot) = \boldsymbol{\phi}(\cdot)^\top \boldsymbol{w}, \tag{6.15}$$

[6] This paragraph and Appendix 6.F are added in the Fall of 2014.

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.[7] 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:

$$\boldsymbol{\Phi} = \Big(\boldsymbol{\phi}(Z_1), \ldots, \boldsymbol{\phi}(Z_n)\Big)^\top, \; \boldsymbol{\Phi}' = \Big(\boldsymbol{\phi}(Z_1'), \ldots, \boldsymbol{\phi}(Z_n')\Big)^\top, \; \boldsymbol{r} = \Big(R_1, \ldots, R_n\Big)^\top, \tag{6.16}$$

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 result.

Proposition 6.1 (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 (6.15) and (6.16), we can rewrite (6.11)-(6.12) as

$$\boldsymbol{u}^*(\boldsymbol{w}) = \operatorname*{argmin}_{\boldsymbol{u} \in \mathbb{R}^p} \left\{ \frac{1}{n} \big[\boldsymbol{\Phi}\boldsymbol{u} - (\boldsymbol{r} + \gamma \boldsymbol{\Phi}' \boldsymbol{w})\big]^\top \big[\boldsymbol{\Phi}\boldsymbol{u} - (\boldsymbol{r} + \gamma \boldsymbol{\Phi}' \boldsymbol{w})\big] + \lambda_{h,n} \boldsymbol{u}^\top \boldsymbol{\Psi} \boldsymbol{u} \right\}, \tag{6.17}$$

$$\begin{aligned} \boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^p} \Big\{ & \frac{1}{n} \big[\boldsymbol{\Phi}\boldsymbol{w} - (\boldsymbol{r} + \gamma \boldsymbol{\Phi}' \boldsymbol{w})\big]^\top \big[\boldsymbol{\Phi}\boldsymbol{w} - (\boldsymbol{r} + \gamma \boldsymbol{\Phi}' \boldsymbol{w})\big] - \\ & \frac{1}{n} \big[\boldsymbol{\Phi}\boldsymbol{u}^*(\boldsymbol{w}) - (\boldsymbol{r} + \gamma \boldsymbol{\Phi}' \boldsymbol{w})\big]^\top \big[\boldsymbol{\Phi}\boldsymbol{u}^*(\boldsymbol{w}) - (\boldsymbol{r} + \gamma \boldsymbol{\Phi}' \boldsymbol{w})\big] + \lambda_{Q,n} \boldsymbol{w}^\top \boldsymbol{\Psi} \boldsymbol{w} \Big\}. \end{aligned} \tag{6.18}$$

Taking the derivative of (6.17) w.r.t. $\boldsymbol{u}$ and equating it to zero, we obtain $\boldsymbol{u}^*$ as a function of $\boldsymbol{w}$:

$$\boldsymbol{u}^*(\boldsymbol{w}) = \Big(\boldsymbol{\Phi}^\top \boldsymbol{\Phi} + n\lambda_{h,n} \boldsymbol{\Psi}\Big)^{-1} \boldsymbol{\Phi}^\top (\boldsymbol{r} + \gamma \boldsymbol{\Phi}' \boldsymbol{w}) = \boldsymbol{A}(\boldsymbol{r} + \gamma \boldsymbol{\Phi}' \boldsymbol{w}). \tag{6.19}$$

Plug $\boldsymbol{u}^*(\boldsymbol{w})$ from (6.19) into (6.18), 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.

Proposition 6.2 (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 (6.15) and (6.16), we can rewrite (6.13)-(6.14) as

$$\boldsymbol{u}^*(\boldsymbol{w}) = \operatorname*{argmin}_{\boldsymbol{u} \in \mathbb{R}^p} \left\{ \frac{1}{n} \big[\boldsymbol{\Phi}\boldsymbol{u} - (\boldsymbol{r} + \gamma \boldsymbol{\Phi}' \boldsymbol{w})\big]^\top \big[\boldsymbol{\Phi}\boldsymbol{u} - (\boldsymbol{r} + \gamma \boldsymbol{\Phi}' \boldsymbol{w})\big] + \lambda_{h,n} \boldsymbol{u}^\top \boldsymbol{\Psi} \boldsymbol{u} \right\}, \tag{6.20}$$

$$\boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^p} \left\{ \big[\boldsymbol{\Phi}\boldsymbol{w} - \boldsymbol{\Phi}\boldsymbol{u}^*(\boldsymbol{w})\big]^\top \big[\boldsymbol{\Phi}\boldsymbol{w} - \boldsymbol{\Phi}\boldsymbol{u}^*(\boldsymbol{w})\big] + \lambda_{Q,n} \boldsymbol{w}^\top \boldsymbol{\Psi} \boldsymbol{w} \right\}. \tag{6.21}$$

Similar to the parametric REG-BRM, we solve (6.20) and obtain $\boldsymbol{u}^*(\boldsymbol{w})$ which is the same as (6.19). If we plug this $\boldsymbol{u}^*(\boldsymbol{w})$ into (6.21), take derivate w.r.t. $\boldsymbol{w}$, and find the minimizer, the parameter vector $\boldsymbol{w}^*$ will be as announced. □

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

RKHS Formulation for REG-BRM and REG-LSTD

A flexible and powerful possibility for choosing the function space $\mathcal{F}^{|\mathcal{A}|}$ is to work with a reproducing kernel Hilbert space $\mathcal{H} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ defined by a positive definite kernel $\kappa : (\mathcal{X} \times \mathcal{A}) \times (\mathcal{X} \times \mathcal{A}) \to \mathbb{R}$, and to use the corresponding RKHS norm $\|\cdot\|_{\mathcal{H}}^2$ as the regularizer $J^2(\cdot)$. REG-BRM with an RKHS function space $\mathcal{F}^{|\mathcal{A}|} = \mathcal{H}$ would be

$$\hat{h}_n(\cdot; Q) = \operatorname*{argmin}_{h \in \mathcal{F}^{|\mathcal{A}|}[=\mathcal{H}]} \Big[\Big\| h - \hat{T}^{\pi_i} Q \Big\|_{\mathcal{D}_n}^2 + \lambda_{h,n} \|h\|_{\mathcal{H}}^2 \Big], \tag{6.22}$$

$$\hat{Q}^{(k)} = \operatorname*{argmin}_{Q \in \mathcal{F}^{|\mathcal{A}|}[=\mathcal{H}]} \Big[\Big\| Q - \hat{T}^{\pi_i} Q \Big\|_{\mathcal{D}_n}^2 - \Big\| \hat{h}_n(\cdot; Q) - \hat{T}^{\pi_i} Q \Big\|_{\mathcal{D}_n}^2 + \lambda_{Q,n} \|Q\|_{\mathcal{H}}^2 \Big], \tag{6.23}$$

and the coupled optimization problems for REG-LSTD are

$$\hat{h}_n(\cdot; Q) = \operatorname*{argmin}_{h \in \mathcal{F}^{|\mathcal{A}|}[=\mathcal{H}]} \Big[\Big\| h - \hat{T}^{\pi_i} Q \Big\|_{\mathcal{D}_n}^2 + \lambda_{h,n} \|h\|_{\mathcal{H}}^2 \Big], \tag{6.24}$$

$$\hat{Q}^{(k)} = \operatorname*{argmin}_{Q \in \mathcal{F}^{|\mathcal{A}|}[=\mathcal{H}]} \Big[\Big\| Q - \hat{h}_n(\cdot; Q) \Big\|_{\mathcal{D}_n}^2 + \lambda_{Q,n} \|Q\|_{\mathcal{H}}^2 \Big]. \tag{6.25}$$

We can solve these coupled optimization problems by the application of the Generalized Representer theorem for RKHS [Schölkopf et al., 2001] (quoted as Theorem B.1 in Appendix B.1.1). 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 6.3. *Let $\tilde{Z}$ be a vector defined as $\tilde{Z} = (Z_1, \dots, Z_n, Z'_1, \dots, Z'_n)^\top$. Then the optimizer $\hat{Q} \in \mathcal{H}$ of (6.22)-(6.23) can be written as $\hat{Q}(\cdot) = \sum_{i=1}^{2n} \tilde{\alpha}_i \kappa(\tilde{Z}_i, \cdot)$ for some values of $\tilde{\alpha} \in \mathbb{R}^{2n}$. The same holds for the solution to (6.24)-(6.25). 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 6.A. □

6.4 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.[8]

Our study has two main parts. First we analyze the policy evaluation error of REG-LSTD in Section 6.4.1. We suppose that given any policy π, we obtain $\hat{Q}$ by solving (6.13)-(6.14)

[8] This section has been revised in the Fall of 2014. The noticeable changes are that most proofs and auxiliary results have been deferred to appendices, and some discussions have been slightly revised or expanded. On the technical side, the proof of Theorem 6.8 (Theorem 6.6 in the original 2011 dissertation) is modified. The upper bounds still have the same general form, but as a result of this change the multiplicative terms are tighter. The conclusions we get from the results remain the same.

with π_k in these equations being replaced by π. Theorem 6.4 provides an upper bound on the Bellman error $\|\hat{Q} - T^{\pi}\hat{Q}\|_{\nu}$. Next in Section 6.4.2, we show how the Bellman errors of the policy evaluation procedure propagate through the API procedure (Theorem 6.5). The main result of this chapter, which is an upper bound on the performance loss $\|Q^* - Q^{\pi_K}\|_{1,\rho}$, is stated as Theorem 6.6 in Section 6.4.3, followed by the discussion of the result.

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 A13 (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., $|R_t| \leq R_{\max}$ $(t = 1, 2, \dots)$ and $\|r\|_{\infty} \leq R_{\max}$.

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 A13 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, 2010; 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 A16.

Assumption A14 (Sampling) At iteration k of REG-LSPI (for $k = 0, \dots, K-1$), n fresh 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 A14 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, (X_t, A_t) gradually "forgets" its past. One way to quantify this forgetting is through mixing processes. For these processes, tools such as *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, in Chapter 4 for analyzing the regularized regression problem, and in Chapter 7 in the context of model selection for RL problems.

We 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 indeed in the off-policy sampling setting. Moreover, changing ν at each iteration based on the previous iterations is a possibility with practical benefits and theoretical justifications [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 datasets in all iterations. We comment on it later after the proof of Theorem 6.4 in Appendix 6.B.

Assumption A15 (Regularizer) Define two regularizer 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.[9] For all $Q \in \mathcal{F}^{|\mathcal{A}|}$ and $a \in \mathcal{A}$, we have $J(Q(\cdot, a)) \leq J(Q)$.

[9] Note that here we are slightly abusing notation 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 a specific expression the identity of the regularizer should always be clear from the context.

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 only is used in the analysis. This assumption imposes some mild restrictions on these regularizer functionals. The condition that the regularizers be pseudo-norms is satisfied by many commonly-used regularizers such as the Sobolev norms, the RKHS norm, and the l_2-regularizer defined in Section 6.3.1 with a positive 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 A16 (Capacity of Function Space) For $R > 0$, let $\mathcal{F}_R = \{f \in \mathcal{F} : J(f) \leq R\}$. There exists 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 bound. The metric entropy is precisely defined in Appendix B.2, 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 $\log \mathcal{N}_\infty(u, \mathcal{F}_R) \leq \log \mathcal{N}_\infty(u, \mathcal{F}) = |\mathcal{X}| \log(\frac{Q_{\max}}{u})$. This shows that the metric entropy for problems with finite state spaces grows much slower than what we consider here. Assumption A16 is suitable for large function spaces and is indeed satisfied for 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 A15 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 A15 and A16.

The metric entropy here is defined w.r.t. the supremum norm. All proofs, except that of Lemma 6.14, 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 6.14.

We can now remark on the requirement that $\mathcal{X}$ is compact (Assumption A13). 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 A13 and implicitly let Assumption A16 satisfy it, but we preferred to be explicit about it at the cost of a bit of redundancy in our set of assumptions.

Assumption A17 (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 A17 requires all the functions in $\mathcal{F}^{|\mathcal{A}|}$ to be bounded so that the solutions of optimization problems (6.13)-(6.14) 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 – quoted in Appendix B.4).

Assumption A18 (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 A19 (Expansion of Smoothness) For all $Q \in \mathcal{F}^{|\mathcal{A}|}$, there exists constants $0 \leq L_R, L_P < \infty$, dependent only on the MDP and $\mathcal{F}^{|\mathcal{A}|}$, such that for any policy π,

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

We require that the complexity of $T^\pi Q$ 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 can still be represented well with a function from $\mathcal{F}^{|\mathcal{A}|}$. Proposition 6.16 in Appendix 6.G presents the conditions that for the so-called *convolutional* MDPs, Assumption A19 holds. Briefly speaking, the conditions are 1) the transition probability kernel should have a finite gain (in 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.

6.4.1 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 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 $\mathcal{D}_n$ and $\hat{Q}$, and $\hat{h}_n(Q) = \hat{h}_n(\cdot; Q)$ refer to the solution to (6.13)-(6.14) 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 6.4 (Policy Evaluation)**.** *For any fixed policy π, let $\hat{Q}$ be the solution to the optimization problem (6.13)-(6.14) 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 A13–A19 hold, there exists $c(\delta) > 0$ such that for any $n \in \mathbb{N}$ and $0 < \delta < 1$, we have

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

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 6.4, which is proven in Appendix 6.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.[10] It shows that if the parameters of the REG-LSTD algorithm is selected properly, one may achieve the sample complexity upper bound of $O(n^{-1/(1+\alpha)})$. This upper bound, as we discuss after stating Theorem 6.6, is optimal for the policy evaluation task. One may note that the proper selection of the parameters requires the knowledge of some unknown quantities such as α and $J(Q^\pi)$. 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 6.5.

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 RHS . One can see that for a fixed Q, the optimization problem (6.13) 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 (6.14) 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 (6.13) is not fixed, but actually 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 6.6 is depicted in Figure 6.2 in Appendix 6.B.

We mentioned earlier that one can actually reuse a single dataset 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 6.4. 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 $O((\frac{n}{K})^{-\frac{1}{2(1+\alpha)}})$. For finite K, or slowly growing K, this is essentially the same as $O(n^{-\frac{1}{2(1+\alpha)}})$.

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

6.4.2 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 **B**ellman **R**esidual (BR) of the k^{th} action-value function

$$\varepsilon_k^{\text{BR}} = \hat{Q}^{(k)} - T^{\pi_k}\hat{Q}^{(k)}. \tag{6.26}$$

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 Theorems 6.5, which is the same as Theorem 3.2 in Chapter 3.

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] and Chapter 3 of this thesis.

Definition 6.1 (Expected Concentrability of the Future State-Action Distribution)**.** *Given $\rho, \nu \in \mathcal{M}(\mathcal{X} \times \mathcal{A})$, $m \geq 0$, and an arbitrary sequence of stationary policies $(\pi_m)_{m \geq 1}$, let $\rho P^{\pi_1} P^{\pi_2} \cdots P^{\pi_m} \in \mathcal{M}(\mathcal{X} \times \mathcal{A})$ 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:

$$a_k = \frac{(1-\gamma)\gamma^{K-k-1}}{1-\gamma^{K+1}}. \qquad (0 \leq k < K) \tag{6.27}$$

Theorem 6.5 (Error Propagation for API – Theorem 3.2 in Chapter 3)**.** *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 (6.26), we have*

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

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

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

6.4.3 Performance Loss of REG-LSPI

In this section, we use the error propagation result (Theorem 6.5 in Section 6.4.2) together with the upper bound on the policy evaluation error (Theorem 6.4 in Section 6.4.1) 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 6.6. *Let $(\hat{Q}^{(k)})_{k=0}^{K-1}$ be the solutions of the optimization problem (6.13)-(6.14) 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 A13–A17 hold; Assumptions A18 and A19 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 $C_{LSTD}(\delta, K; \rho, \nu)$ such that for any $n \in \mathbb{N}$ and $0 < \delta < 1$, we have

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

with probability at least $1 - \delta$.

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

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

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 $k = 0, \dots, K-1$, invoke Theorem 6.4 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

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

which holds with probability at least $1 - \frac{\delta}{K}$. Here $c(\cdot)$ is defined as in Theorem 6.4. For any $r \in [0,1]$, we have

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

where we used the definition of a_k (6.27). We then apply Theorem 6.5 with the choice of $p = 1$ to get that with probability at least $1 - \delta$, we have

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

Here

$$C_{\mathrm{LSTD}}(\rho, \nu; K) =$$

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

□

Theorem 6.6 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 effect of number of samples in the upper bound is $O(n^{-\frac{1}{2(1+\alpha)}})$. This upper bound is notable because it is the minimax rate for the regression problem when the regression function belongs to a function space $\mathcal{F}$ with a packing entropy in the same form as in the upper bound of Assumption A16 [Yang and Barron, 1999]. Since the regression problem is a subset of the policy evaluation subtask of RL/Planning problem, the minimax rate for regression is also a minimax rate for the policy evaluation in RL/Planning problems. Nevertheless, the optimality of the error bound for the policy evaluation task does not necessarily imply that the 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 goal of RL, even though the estimate of the action-value function is inaccurate. Refer to Farahmand [2011] for more discussion.

The term C_{LSTD} has two main components. The first is $C_{\text{PI},\rho,\nu}(\cdot;r)$ that 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 in Chapter 3, but briefly speaking it indicates that ν and ρ affect the performance through a weighted summation of $c_{\text{PI}_1,\rho,\nu}$ (Definition 6.1). The concentrability coefficients $c_{\text{PI}_1,\rho,\nu}$ is defined as the squared 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 6.5 actually provides a stronger result than what is reported in Theorem 6.6: the effect of errors at earlier iterations on the performance loss is geometrically decayed. So one may potentially use a smaller 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 $\pi_0, \pi_1, \dots, \pi_{K-1}$ generated by the REG-LSPI procedure. One might expect that as the sequence $\hat{Q}^{(k)}$ converge to a neighbourhood of Q^*, the value of $J(Q^{\pi_{k+1}})$ gets close to $J(Q^*)$. We postpone the analysis of this finer structure of the problem to future work.

Comparison with similar statistical guarantees

Theorem 6.6 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], where the statistical properties of LASSO-TD are analyzed. Although these results address different algorithms, comparing them can still be insightful.[11]

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

[11] The material of this section is extended compared to the originally submitted dissertation to include [Ghavamzadeh et al., 2011; Lazaric et al., 2012; Ávila Pires and Szepesvári, 2012].

our $C_{\mathrm{PI},\rho,\nu}(K;r)$. In contrast, our simplified upper bound is $C_{\mathrm{LSTD}}(\delta)n^{-\frac{1}{2(1+\alpha)}}$, in which $C_{\mathrm{LSTD}}(\delta)$ can roughly be factored into $C^{\frac{1}{2}}_{\mathrm{PI},\rho,\nu}(K;r)\, C_1(J(Q^\pi), L_R, L_P)\, \sqrt{\ln(K/\delta)}$.

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 A16. 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 $n^{-1/2}$. 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]. It is also notable that for problems with finite state space, we have $\log \mathcal{N}_\infty(u, \mathcal{F}_R) \leq |\mathcal{X}| \log(\frac{Q_{\max}}{u})$, which leads to $n^{-1/2}$ error upper bound (disregarding the logarithmic terms).

The other difference is in the definition of concentrability coefficients ($C_{\mathrm{PI},\rho,\nu}(K)$ vs. $C_{\rho,\nu}$). In Definition 6.1, 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 also be used to improve the result of Antos et al. [2008b] too, refer to Chapter 3.

Lazaric et al. [2012] analyzed 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 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. Their formulation has some similarities to REG-LSTD (6.13)-(6.14). Beside the difference between the choice of the class of function spaces with this work, another algorithmic difference is that there is no regularization in their projection step (6.13). 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. [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, however 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-LSPI with a linear function space (similar to Section 6.3.1) with $J^2(h) = \|u\|_1$ and $J^2(Q) = \|w\|_1$, the current analysis leads to $O(\|w^*\|_1^{1/2} n^{-1/4})$ error upper bound with a logarithmic dependence on p. This result might be obtained using Corollary 5 of Zhang [2002] as Assumption A16. 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 (6.13)-(6.14). 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.

6.5 Conclusion and Future Work

In this work we introduced two regularization-based API algorithms to solve RL/Planning problems with large state spaces. The core of these algorithms are novel policy evaluation methods, namely REG-BRM and REG-LSTD, that estimate the action-value functions by solving two coupled optimization problems with regularized objective functions. Our formulation was general and could incorporate many types of function spaces and regularizers.

We showed how these algorithms can be implemented efficiently when the function space is either the span of a finite number of basis functions (parametric model) or an RKHS (nonparametric model). The RKHS formulation has some advantages such as the generality to work with different input domains and the ease of choosing/changing the kernel function and consequently the function space. This flexibility is a key ingredient for an adaptive algorithm.

Afterwards, we focused on the statistical properties of REG-LSPI and provided an error upper bound on the performance loss of the resulting policy (Theorem 6.6). The error bound showed the role of the sample size, complexity of function space (quantified by its metric entropy in Assumption A16), and the intrinsic properties of MDP such as the behavior of concentrability coefficients and the smoothness-expansion property of the Bellman operator (Assumption A19). The result showed that the dependence on the sample size for the task of policy evaluation is optimal.

To our best knowledge this (and its conference version [Farahmand et al., 2009b]) alongside our other work on Regularized Fitted Q-Iteration [Farahmand et al., 2008, 2009a] and Chapter 5 are the first work that address the finite-sample performance of a regularized RL algorithm. Nevertheless, there have been a few other work that also used regularization for RL/Planning problems without thoroughly analyzing them. We briefly mention them in the what follows.

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 sparsification techniques. 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 LSPI. They used sparsification to provide basis functions for the LSTD procedure. Although sparsification controls the complexity of the estimate, to our best knowledge its effect on the generalization error is not well-understood. Sparsification is fundamentally different from our approach. In our method, the empirical error and the regularization term jointly determine the solution. In sparsification methods, however, one selects a subset of data points based on some criteria and then use them as basis functions. Kolter and Ng [2009] formulated an l_1-regularization extension of LSTD and provided LARS-like algorithm [Efron et al., 2004]

to efficiently 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], which we discussed earlier. 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 estimators, or both. Their result was obtained by considering two separate regularized regression problems: one with the reward function and the other with $\kappa(X, X')$ as the regression function. Thus it is different from our formulation that is stated as a coupled optimization problem in an RKHS.

This work is a step forward to understand regularization-based algorithms for solving RL/Planning problems. We briefly comment on several possibilities for future studies.

Computational Considerations. Devising a computationally efficient implementation of REG-LSPI/BRM is important to ensure that it is a practical algorithm for real-world problems. The 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 approach is to use stochastic gradient methods to approximately solve the corresponding optimization problem. This is especially appealing in the light of results such as Bottou and Bousquet [2008] who show that given a fixed amount of computation time, the generalization error resulting from learning with stochastic gradient methods as the optimizer might be less than that of gradient-descent methods. Refer to Section 5.6 for more detailed discussion.

Other Regularizers. Our formulation of REG-BRM and REG-LSTD is general for many classes of regularizers J. Our statistical analysis is also valid whenever the function space satisfies the required metric entropy assumption (Assumption A16). Our closed-form solutions of Section 6.3.1, however, are specially tailored to RKHS with its inner-product norm and to parametric models with the l_2-norm as the regularizer. One may indeed think of other types of regularizers, such as total variation norm [Mammen and van de Geer, 1997] or l_1-norm.

The l_1-regularization is a viable possibility and can be used to exploit sparsity of the action-value function similar to the way it has been used in regression [Tibshirani, 1996]. Nevertheless, the use of the l_1-regularization for LSTD/BRM is not computationally straightforward. The reason is that if we want to use the l_1-norm in (6.11)-(6.12) or (6.13)-(6.14) as the regularizer J, we do not have a closed-form solution for the coupled optimization problems anymore. This prevents us from plugging $\hat{h}_n(\cdot; Q)$ directly into the second optimization problem.

One idea is to solve these two optimization problems concurrently by a gradient-descent method, and plug the most recent solution of $\hat{h}_n(\cdot; Q)$ into the second optimization problem. One should show that this procedure has a unique stable fixed point. By using a two-time-scale gradient-descent algorithm, *singular perturbation theory* [Khalil, 2001, Chapter 11] might provide a way to prove the convergence to a close neighborhood of the original fixed-point solution. This claim requires further investigations.

Continuous Action Space. A practically important question is how to extend REG-LSPI/BRM to deal with continuous action MDPs as well. Again, we face the same difficulties as discussed in Section 5.6.

Influence of the MDP on Smoothness. An open theoretical question is to characterize

the properties of MDP that determine the function space to which the action-value function belongs. A similar question is how the values of L_P and L_R in Assumption A19 are related to the intrinsic properties of the MDP. We partially addressed this question for the convolutional MDPs, but analysis for more general MDPs is remained to be done.

Model Selection. An important issue in the successful application of any RL/Planning algorithm, including REG-LSPI/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 $\lambda_{h,n}$. In Theorem 6.6 we select $\lambda_{h,n} = \lambda_{Q,n} = [\frac{1}{n\, J^2(Q^\pi)}]^{\frac{1}{1+\alpha}}$. This choice, however, requires the knowledge of $J(Q^\pi)$ that is not available. In addition, we have assumed that Q^π is in $\mathcal{F}^{|\mathcal{A}|}$ and the application of T^π on some $Q \in \mathcal{F}^{|\mathcal{A}|}$ is well-behaving, i.e., the roughness of $T^\pi Q$ measured according to the natural norm of the space, $J(T^\pi Q)$, is not much larger than $J(Q)$. In other words, one must select a function space $\mathcal{F}^{|\mathcal{A}|}$ that "matches" with the MDP. In Chapter 7, we address the question of parameter selection in the RL/Planning context and introduce BErMin, a complexity-regularization-based model selection algorithm, and analyze its properties.

Other Technical Questions. A technical question that has not yet been addressed is how to extend our results from the i.i.d. process to more general mixing processes [Doukhan, 1994; Yu, 1994]. A possible approach is to use independent block technique – similar to Chapter 4. We postpone this extension to future research. Another technical question is whether it is possible to relax the supremum norm requirement in metric entropy assumption (Assumption A16).

Appendices

Proofs and Auxiliary Results

In these appendices, we first prove Theorem 6.3, which provides the closed-form solutions for REG-LSTD and REG-BRM when the function space is an RKHS (Appendix 6.A). We then attend to the proof of Theorem 6.4 (Policy Evaluation error for REG-LSTD). The main body of proof for Theorem 6.4 is in Appendix 6.B. To increase the readability and flow, the proofs of some of the auxiliary and more technical results are postponed to Appendices 6.C, 6.D, and 6.E.

More specifically, we prove an extension of Theorem 21.1 of Györfi et al. [2002] in Appendix 6.C (Lemma 6.7). We present a modified version of Theorem 10.2 of van de Geer [2000] in Appendix 6.D. We then provide a covering number result in Appendix 6.E (Lemma 6.12). The reason we require these results will be clear in Appendix 6.B.

We explain why for large function spaces we need to use regularizers in both optimization problems (6.13) and (6.14) of REG-LSTD in Appendix 6.F. Finally, we introduce convolutional MDPs as an instance of problems that satisfy Assumption A19 (Appendix 6.G).

6.A Proof of Theorem 6.3 (Closed-form solutions for RKHS formulation of REG-LSTD/BRM)

Proof. **REG-BRM:** First, notice that the optimization problem (6.23) can be written in the form $c_n(Q) + \lambda_{Q,n} \|Q\|^2_{\mathcal{H}} \overset{Q}{\to} \min!$ with an appropriately defined functional c_n.[12] 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 (6.22), calling for the Representer theorem once

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

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

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

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 $(\hat{T}^{\pi_k} Q)$ is viewed as the n-dimensional vector

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

Thus, $\boldsymbol{\beta}^*$ depends on Q only through $Q(Z_1'), \dots, Q(Z_n')$. Plugging this solution into (6.23), 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 $\tilde{\boldsymbol{\alpha}} = (\alpha_1, \dots, \alpha_n, \alpha_1', \dots, \alpha_n')^\top$, using the reproducing kernel property of K we get the optimization problem

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

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 $Z_1, Z_1', \dots, Z_n, Z_n'$ is exactly the same as the proof of REG-BRM. Thus, using the Representer theorem, the minimizer of (6.25) 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 $\tilde{\boldsymbol{\alpha}} = (\alpha_1, \dots, \alpha_n, \alpha_1', \dots, \alpha_n')^\top$, using the reproducing kernel property of K we get the optimization problem

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

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

6.B Proof of Theorem 6.4 (Statistical guarantee for REG-LSTD)

The goal of Theorem 6.4 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 (6.13) and (6.14). 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 right-hand side. Recall from the discussion after Theorem 6.4 that the analysis is more complicated than the 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 6.4 is proven using Lemma 6.7, which upper bounds $\|\hat{h}_n(\cdot; \hat{Q}) - T^\pi \hat{Q}\|_\nu$, and Lemma 6.11, which upper bounds $\|\hat{Q} - \hat{h}_n(\cdot; \hat{Q})\|_\nu$. We also require to relate the smoothness

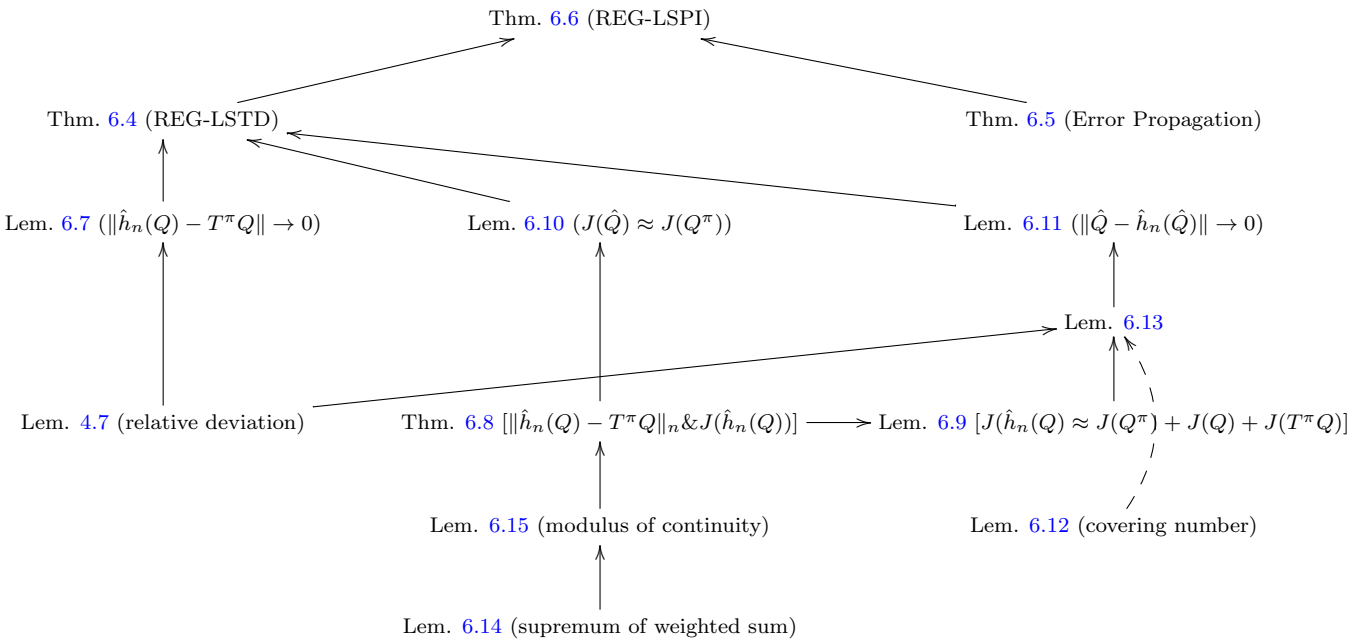

Figure 6.2: Dependencies of results used to prove the statistical guarantee for REG-LSPI (Theorem 6.6).

$J(\hat{Q})$ to the smoothness $J(Q^\pi)$. Lemma 6.10 specifies this relation. The proof of these lemmas themselves require further developments, which will be discussed when we encounter them. Figure 6.2 shows the dependencies between all results that lead to the proof of Theorem 6.4 and consequently Theorem 6.6.

The following lemma controls the error behavior resulting from the optimization problem (6.13). 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. Firstly, it holds uniformly over $T^\pi Q$ (as opposed to a fixed function $T^\pi Q$); secondly, it holds for function spaces that satisfy a general metric entropy condition (as opposed to the special case of Sobolev spaces).

Lemma 6.7 (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 (6.13). Under Assumptions A13–A17, there exist finite constants $c_1, c_2 > 0$ such that for any $n \in \mathbb{N}$ and $0 < \delta < 1$, we have*

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

with probability at least $1 - \delta$.

Proof. See Appendix 6.C. □

When we use this lemma to prove Theorem 6.4, the action-value function Q that appears in the bound is the result of the optimization problems defined in (6.14), and so is random. Lemma 6.10, which we will prove later, provides a deterministic upper bound for 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 6.7. In particular, we need an upper bound on the empirical error of the regularized regression estimator $\hat{h}_n(\cdot; Q)$

(cf. (6.28)). 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 6.8 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 dataset $\mathcal{D}_n$ 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:

$$\hat{h}_n(\cdot;Q) = \operatorname*{argmin}_{h \in \mathcal{F}^{|\mathcal{A}|}} \left[\|h - [R_i + \gamma Q(X_i', \pi(X_i'))]\|_n^2 + \lambda_{h,n} J^2(h) \right], \tag{6.28}$$

which is the same as (6.13) with π replacing π_k.

Theorem 6.8 (Empirical error and smoothness of $\hat{h}_n(\cdot;Q)$)**.** *For a random function Q, let $\hat{h}_n(\cdot,Q)$ be defined according to (6.28). Suppose that Assumptions A13-A17 hold. Then, there exist constants $c_1, c_2 > 0$, such that for any $n \in \mathbb{N}$ and $0 < \delta < 1$, we have*

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

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

with probability at least $1 - \delta$.

Proof. See Appendix 6.D. □

The following lemma, which is an immediate corollary of Theorem 6.8, 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 6.13, which itself is used in the proof of Lemma 6.11.

Lemma 6.9 (Smoothness of $\hat{h}_n(\cdot;Q)$)**.** *For a random Q, let $\hat{h}_n(\cdot;Q)$ be the solution to the optimization problem (6.13) with the choice of regularization coefficient*

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

Let Assumptions A13–A17 hold. Then there exits a finite constant $c > 0$, depending on Q_{max}, such that for any $n \in \mathbb{N}$ and $0 < \delta < 1$,

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

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 6.8 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 $0 < \delta < 1$,

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

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 6.4.

Lemma 6.10 (Smoothness of $\hat{Q}$)**.** *Let Assumptions A13–A18 hold, and let $\hat{Q}$ be the solution to (6.14) 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

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

with probability at least $1-\delta$.

Proof. By the optimizer property of $\hat{Q}$, we have

$$\lambda_{Q,n} J^2(\hat{Q}) \leq \left\| \hat{Q} - \hat{h}_n(\cdot; \hat{Q}) \right\|_{\mathcal{D}_n}^2 + \lambda_{Q,n} J^2(\hat{Q}) \leq \left\| Q^\pi - \hat{h}_n(\cdot; Q^\pi) \right\|_{\mathcal{D}_n}^2 + \lambda_{Q,n} J^2(Q^\pi). \tag{6.29}$$

Since by Assumption A18, $Q^\pi = T^\pi Q^\pi \in \mathcal{F}^{|\mathcal{A}|}$, Theorem 6.8 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 $0 < \delta < e^{-1} \approx 0.3679$,

$$\left\| Q^\pi - \hat{h}_n(\cdot; Q^\pi) \right\|_{\mathcal{D}_n}^2 \leq c_1 \left(1 \vee Q_{\max}^{2(1+\alpha)}\right) \frac{J^{\frac{2\alpha}{1+\alpha}}(Q^\pi) \ln(1/\delta)}{n^{\frac{1}{1+\alpha}}} \tag{6.30}$$

holds with probability at least $1-\delta$. Chaining inequalities (6.29) and (6.30) finishes the proof. □

The other main ingredient of the proof of Theorem 6.4 is an upper bound to $\|\hat{Q} - \hat{h}_n(\cdot; \hat{Q})\|_\nu$, which is closely related to the optimization problem (6.14). This task is done by Lemma 6.11. In the proof of this lemma, we call Lemma 6.13, which shall be stated and proven right after this result.

Lemma 6.11 (Convergence of $\|\hat{Q} - \hat{h}_n(\cdot; \hat{Q})\|_\nu$)**.** *Let $\hat{Q}$ be the solution to the set of coupled optimization problems (6.13)-(6.14). Suppose that Assumptions A13–A19 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

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

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

$$\begin{aligned}\frac{1}{2}I_{1,n} &= \left\|\hat{Q} - \hat{h}_n(\cdot;\hat{Q})\right\|_{\mathcal{D}_n}^2 + \lambda_{Q,n}J^2(\hat{Q}),\\ I_{2,n} &= \left\|\hat{Q} - \hat{h}_n(\cdot;\hat{Q})\right\|_\nu^2 - I_{1,n}.\end{aligned} \tag{6.31}$$

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

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

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 6.8. 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

$$\frac{1}{2}I_{1,n} \leq \lambda_{Q,n}J^2(Q^\pi) + c_1\frac{J^{\frac{2\alpha}{1+\alpha}}(Q^\pi)\ln(1/\delta_1)}{n^{\frac{1}{1+\alpha}}}, \tag{6.32}$$

with probability at least $1 - \delta_1$.
$I_{2,n}$: With our choice of $\lambda_{Q,n}$ and $\lambda_{h,n}$, Lemma 6.13, 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

$$I_{2,n} \leq c_2\frac{L_R^{\frac{2\alpha}{1+\alpha}} + [J(Q^\pi)]^{\frac{2\alpha}{1+\alpha}}[\ln(1/\delta_2)]^{\frac{\alpha}{1+\alpha}}}{n^{\frac{1}{1+\alpha}}} + c_3\frac{(1+\gamma^2L_P^2)^\alpha}{n\,\lambda_{Q,n}^\alpha} + c_4\frac{\ln(1/\delta_2)}{n}, \tag{6.33}$$

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

$$\frac{1}{n\,\lambda_{Q,n}^\alpha} = \frac{[J(Q^\pi)]^{\frac{2\alpha}{1+\alpha}}}{n^{\frac{1}{1+\alpha}}} \leq \frac{[J(Q^\pi)]^{\frac{2\alpha}{1+\alpha}}}{n^{\frac{1}{1+\alpha}}}\ln(1/\delta_2). \tag{6.34}$$

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 (6.32), (6.33), and (6.34) 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

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

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

To upper bound $I_{2,n}$, defined in (6.31), we simultaneously apply the peeling device (cf. Section 5.3 of van de Geer 2000; also Appendix B.3) 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 4.7 in Appendix 4.A). We also require a covering number result, which is stated as Lemma 6.12. The final result of this procedure is a tight upper bound on $I_{2,n}$, as stated in Lemma 6.13.

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

$$\begin{aligned}\mathcal{F}_\sigma &\triangleq \{f : f \in \mathcal{F}, J^2(f) \leq \sigma\},\\ \mathcal{F}_\sigma^{|\mathcal{A}|} &\triangleq \{f : f \in \mathcal{F}^{|\mathcal{A}|}, J^2(f) \leq \sigma\}.\end{aligned}$$

Let

$$g_{Q,h}(x,a) \triangleq \sum_{j=1}^{|\mathcal{A}|} \mathbb{I}_{\{a=a_j\}} \left[Q_j(x) - h_j(x)\right]^2. \tag{6.35}$$

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 $J(Q) \leq \sigma_1$ and $J(h) \leq \sigma_2$, i.e.,

$$G_{\sigma_1,\sigma_2} \triangleq \{g_{Q,h} : \mathbb{R}^d \times \mathcal{A} \to \mathbb{R}; Q \in \mathcal{F}_{\sigma_1}^{|\mathcal{A}|}, h \in \mathcal{F}_{\sigma_2}^{|\mathcal{A}|}\}. \tag{6.36}$$

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

Lemma 6.12 (Covering Number)**.** *Let Assumptions A15, A16, and A17 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 $((x_1,a_1),\ldots,(x_n,a_n)) \in \mathcal{X} \times \mathcal{A}$, the empirical covering number of the class of functions G_{σ_1,σ_2} defined in (6.36) w.r.t. the empirical norm $\|\cdot\|_{2,z_{1:n}}$ is upper bounded by*

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

Proof. See Appendix 6.E. □

Next we state Lemma 6.13, which provides a high probability upper bound on $I_{2,n}$.

Lemma 6.13. *Let $I_{2,n}$ be defined according to (6.31). Under Assumptions A13–A17 and A19 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 exists 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 $1-\delta$.

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 6.12 and 4.7 to obtain a tight high-probability upper bound on $I_{2,n}$. Based on the definition of $I_{2,n}$ in (6.31) we have

$$\mathbb{P}\{I_{2,n} > t\} = \mathbb{P}\left\{\frac{\mathbb{E}\left[g_{\hat{Q},\hat{h}_n(\cdot;\hat{Q})}(Z)|\mathcal{D}_n\right] - \frac{1}{n}\sum_{i=1}^n g_{\hat{Q},\hat{h}_n(\cdot;\hat{Q})}(Z_i)}{t + 2\lambda_{Q,n} J^2(\hat{Q}) + \mathbb{E}\left[g_{\hat{Q},\hat{h}_n(\cdot;\hat{Q})}(Z)|\mathcal{D}_n\right]} > \frac{1}{2}\right\}. \tag{6.37}$$

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 6.9 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 $\mathbb{P}\{I_{2,n} > t\} = \mathbb{P}\{I_{2,n} > t, \mathcal{A}_0^C\} + \mathbb{P}\{I_{2,n} > t, \mathcal{A}_0\} \leq \delta_1 + \mathbb{P}\{I_{2,n} > t, \mathcal{A}_0\}$, 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 A19 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

$$\sigma_2^l = c\left[2\left(L_R^2 + (1+\gamma^2 L_P^2)\frac{2^l t}{\lambda_{Q,n}}\right) + J^2(Q^\pi)\ln(1/\delta_1)\right]. \tag{6.38}$$

Apply the peeling device on (6.37). Use (6.38) 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

$$\mathbb{P}\{I_{2,n} > t\} = \mathbb{P}\left\{I_{2,n} > t, \mathcal{A}_0^C\right\} + \mathbb{P}\{I_{2,n} > t, \mathcal{A}_0\} \leq \delta_1 +$$
$$\sum_{l=0}^{\infty} \mathbb{P}\left\{\mathcal{A}_0, 2^l t \mathbb{I}_{\{l\neq 0\}} \leq 2\lambda_{Q,n} J^2(\hat{Q}) < 2^{l+1} t, \frac{\mathbb{E}\left[g_{\hat{Q},\hat{h}_n(\cdot;\hat{Q})}(Z)|\mathcal{D}_n\right] - \frac{1}{n}\sum_{i=1}^n g_{\hat{Q},\hat{h}_n(\cdot;\hat{Q})}(Z_i)}{t + 2\lambda_{Q,n} J^2(\hat{Q}) + \mathbb{E}\left[g_{\hat{Q},\hat{h}_n(\cdot;\hat{Q})}(Z)|\mathcal{D}_n\right]} > \frac{1}{2}\right\}$$
$$\leq \delta_1 + \sum_{l=0}^{\infty} \mathbb{P}\left\{\sup_{g_{Q,h}\in G_{\sigma_1^l,\sigma_2^l}} \frac{\mathbb{E}\left[g_{Q,h}(Z)|\mathcal{D}_n\right] - \frac{1}{n}\sum_{i=1}^n g_{Q,h}(Z_i)}{2^l t + \mathbb{E}\left[g_{Q,h}(Z)|\mathcal{D}_n\right]} > \frac{1}{2}\right\}. \tag{6.39}$$

Let us study the behavior of the l^{th} term of the above summation by verifying the conditions of Lemma 4.7 when $\varepsilon = \frac{1}{2}$ and $\eta = 2^l t$.

Condition (A1): Since all functions involved are bounded by $Q_{\max}$, it is easy to see that $|g_{Q,h}(x,a)| \leq \sum_{j=1}^{|\mathcal{A}|} \mathbb{I}_{\{a=a_j\}} \left|[Q_j(x) - h_j(x)]^2\right| \leq 4Q_{\max}^2$. Therefore, K_1, defined in Lemma 4.7, can be set to $K_1 = 4Q_{\max}^2$.

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

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

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

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 $\varepsilon' \geq \frac{1}{8}\eta = \frac{1}{8}2^l t$, the following holds:

$$\frac{\sqrt{n}(\frac{1}{2})(\frac{1}{2})\varepsilon'}{96\sqrt{2}\max\{K_1, 2K_2\}} \geq$$
$$\int_{\frac{(\frac{1}{2})(\frac{1}{2})\varepsilon'}{16\max\{K_1,2K_2\}}}^{\sqrt{\varepsilon'}} \left(\log \mathcal{N}_2\left(u, \{g \in G_{\sigma_1,\sigma_2} : \frac{1}{n}\sum_{i=1}^n g^2(z_i) \leq 16\varepsilon'\}, z_{1:n}\right)\right)^{1/2} du. \tag{6.40}$$

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

$$\log \mathcal{N}_2\left(u, \left\{g \in G_{\sigma_1,\sigma_2} : \frac{1}{n}\sum_{i=1}^n g^2(z_i) \leq 16\varepsilon'\right\}, z_{1:n}\right) \leq \log \mathcal{N}_2\left(u, G_{\sigma_1,\sigma_2}, z_{1:n}\right)$$
$$\leq c\,(\sigma_1^\alpha + \sigma_2^\alpha) u^{-2\alpha}, \tag{6.41}$$

where we used Lemma 6.12 in the second inequality.

Plug (6.41) into (6.40) with the choice of $\sigma_1^l = \frac{2^l t}{\lambda_{Q,n}}$ and $\sigma_2^l = c[2(L_R^2 + (1+\gamma^2 L_P^2)\frac{2^l t}{\lambda_{Q,n}}) + J^2(Q^\pi)\ln(1/\delta_1)]$. Therefore, for some constant $c(Q_{\max})$, the inequality

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

implies (6.40). Because $(a+b)^{\frac{1}{2}} \leq (a^{\frac{1}{2}} + b^{\frac{1}{2}})$ for non-negative a and b, it suffices to verify the following two conditions:
(a) We shall verify that for $\varepsilon' \geq \frac{1}{8} 2^l t$, we have

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

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:

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

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

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

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

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

Let the left-hand side be equal δ and solve for t. Considering all aforementioned conditions, we get that there exists 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 6.4.

Proof of Theorem 6.4. We want to show that $\|\hat{Q} - T^\pi \hat{Q}\|_\nu$ is small. Since (6.13)-(6.14) 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:

$$\left\|\hat{Q} - T^\pi \hat{Q}\right\|_\nu^2 \leq 2\left\|\hat{Q} - \hat{h}_n(\cdot;\hat{Q})\right\|_\nu^2 + 2\left\|\hat{h}_n(\cdot;\hat{Q}) - T^\pi \hat{Q}\right\|_\nu^2. \tag{6.42}$$

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 6.7 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

$$\left\|\hat{h}_n(\cdot;\hat{Q}) - T^\pi \hat{Q}\right\|_\nu^2 \leq \lambda_{h,n}\left(2J^2(\hat{Q}) + 4J^2(T^\pi \hat{Q})\right) + c_1 \frac{1}{n\lambda_{h,n}^\alpha} + c_2 \frac{\ln(3/\delta)}{n}, \tag{6.43}$$

with probability at least $1 - \delta/3$. Note that $T^\pi \hat{Q} \in \mathcal{F}^{|\mathcal{A}|}$ is implied by Assumption A19 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 (6.43) are also random. In order to upper bound them, we use Lemma 6.10 that 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 $n \in \mathbb{N}$,

$$\lambda_{h,n} J^2(\hat{Q}) = \lambda_{Q,n} J^2(\hat{Q}) \leq \lambda_{Q,n} J^2(Q^\pi) + c_3 \frac{J^{\frac{2\alpha}{1+\alpha}}(Q^\pi)}{n^{\frac{1}{1+\alpha}}} \ln(3/\delta) \tag{6.44}$$

holds with probability at least $1 - \delta/3$. Use Assumption A19 to show that

$$\lambda_{h,n} J^2(T^\pi \hat{Q}) \leq 2\lambda_{Q,n} L_R^2 + 2(\gamma L_P)^2 \left(\lambda_{Q,n} J^2(Q^\pi) + c_3 \frac{J^{\frac{2\alpha}{1+\alpha}}(Q^\pi)}{n^{\frac{1}{1+\alpha}}} \ln(3/\delta)\right) \tag{6.45}$$

holds with the same probability. Plugging (6.44) and (6.45) into (6.43) and using the selected schedule for $\lambda_{Q,n}$ and $\lambda_{h,n}$, we get

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

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

$$\left\|\hat{h}_n(\cdot;\hat{Q}) - T^\pi \hat{Q}\right\|_\nu^2 \leq \left[c_4\left[1 + (\gamma L_P)^2\right] J^{\frac{2\alpha}{1+\alpha}}(Q^\pi) \ln(1/\delta) + \frac{8L_R^2}{[J(Q^\pi)]^{\frac{2}{1+\alpha}}}\right] \frac{1}{n^{\frac{1}{1+\alpha}}}, \tag{6.46}$$

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 6.11 states that there exists a constant $c_5 > 0$ such that for any $n \in \mathbb{N}$,

$$\left\|\hat{Q} - \hat{h}_n(\cdot;\cdot;\hat{Q})\right\|_\nu^2 \leq c_5 \frac{(1 + \gamma^2 L_P^2)^\alpha J^{\frac{2\alpha}{1+\alpha}}(Q^\pi) \ln(1/\delta) + L_R^{\frac{2\alpha}{1+\alpha}}}{n^{\frac{1}{1+\alpha}}}, \tag{6.47}$$

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

Thus, inequality (6.42) alongside upper bounds (6.46) and (6.47) indicate that there exist constants $c_6, c_7 > 0$ such that for any $n \in \mathbb{N}$ and $\delta > 0$, we have

$$\left\|\hat{Q} - T^\pi \hat{Q}\right\|_\nu^2 \leq \frac{c_7 \left[1 + (\gamma L_P)^2\right] J^{\frac{2\alpha}{1+\alpha}}(Q^\pi) \ln(1/\delta) + c_8 \left(L_R^{\frac{2\alpha}{1+\alpha}} + \frac{L_R^2}{[J(Q^\pi)]^{\frac{2}{1+\alpha}}}\right)}{n^{\frac{1}{1+\alpha}}}$$

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

A careful study of the proof of Theorem 6.4 and the auxiliary results used in it reveals that one can indeed reuse a single datasets 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 datasets. 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 dataset 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 datasets in earlier iterations. When we assume that all datasets 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 datasets of the earlier iterations. Conditioned on that randomness, the policy π becomes a deterministic function. This is how we presented the statement of Theorem 6.4 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 6.7, 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 effectively means that for the probabilistic upper bound, the randomness of π in $T^\pi Q$ becomes irrelevant as we are providing a uniform over $\mathcal{G}_l$ guarantee. In the proof of this theorem, we also use Lemma 6.11, which itself uses Theorem 6.8 and Lemma 6.13 that have a similar construct.

6.C Proof of Lemma 6.7 (Convergence of $\hat{h}_n(\cdot;Q)$ to $T^\pi Q$)

Proof of Lemma 6.7. 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

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

with

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

By the optimizer property of $\hat{h}_n(\cdot;Q)$, we get

$$I_{1,n} \le 2\Bigg[\frac{1}{n}\sum_{i=1}^{n}\left|T^\pi Q(Z_i) - \hat{T}^\pi Q(Z_i)\right|^2 - \left|T^\pi Q(Z_i) - \hat{T}^\pi Q(Z_i)\right|^2 + \lambda_{h,n}\left(J^2(T^\pi Q) + J^2(Q) + J^2(T^\pi Q)\right)\Bigg]$$
$$= 4\lambda_{h,n}J^2(T^\pi Q) + 2\lambda_{h,n}J^2(Q). \tag{6.48}$$

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

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

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

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

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 smaller 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 4.7 – similar to what we did in the proof of Lemma 6.13.

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

$$t \ge \frac{c_1}{n}, \tag{6.49}$$

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 6.12: 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

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

With the same covering set argument as in the proof of Lemma 6.12, 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 A16 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

$$t \ge \frac{c_2}{n\lambda_{h,n}^{\alpha}}, \tag{6.50}$$

for a constant $c_2 > 0$, which is the function of $Q_{\max}$ and $|\mathcal{A}|$ only. Therefore, Lemma 4.7 indicates that

$$\mathbb{P}\{I_{2,n} > t\} \le 60 \sum_{l=0}^{\infty} \exp\left(-\frac{n(t + 2^l t)(1/4)(1/2)}{128 \times 2304 \times \max\{16Q_{\max}^4, 4Q_{\max}^2\}}\right) \le c_3 \exp(-c_4 n t). \tag{6.51}$$

for some constants $c_3, c_4 > 0$.

Combining (6.48), (6.49), (6.50), and (6.51), we find that there exist constants $c_5, c_6 > 0$ 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 a function of $Q_{\max}$ and $|\mathcal{A}|$ only, and c_6 is a function of $Q_{\max}$. □

6.D Proof of Theorem 6.8 (Empirical error and smoothness of $\hat{h}_n(\cdot; Q)$)

To prove Theorem 6.8, we modify and specialize Lemma 3.2 by van de Geer [2000] to be suitable to our problem. The modification is required because Q in (6.28) 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 $\mathcal{G}' = \left\{ (Q, T^{\pi}Q, \mathbf{1}) : Q \in \mathcal{F}^{|\mathcal{A}|} \right\}$ 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 6.14 (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*

$$\sqrt{n}(\delta - \varepsilon) \ge C\,L \left[\int_{\frac{\varepsilon}{28L}}^{R} [\log \mathcal{N}_\infty(u, \mathcal{G} \times \mathcal{G}')]^{1/2} \mathrm{d}u \ \vee R \right], \tag{6.52}$$

we have

$$\mathbb{P}\left\{ \sup_{(g,g') \in \mathcal{G} \times \mathcal{G}'} \left| \frac{1}{n} \sum_{i=1}^n W_i(g') g(z_i) \right| \ge \delta \right\} \le 2 \exp\left(-\frac{n(\delta - \varepsilon)^2}{6^7 (RL)^2} \right).$$

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. (6.52)) 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 $g_1', g_2' \in \mathcal{G}'$, we have

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^n W_i(g_1') g_1(z_i) - W_i(g_2') g_2(z_i) = \\
&\frac{1}{n} \sum_{i=1}^n (R_i - r(z_i))(g_1(z_i) - g_2(z_i)) \\
&\qquad + \frac{1}{n} \sum_{i=1}^n \gamma \left[(Q_1(X_i', \pi(X_i')) - P^\pi Q_1(z_i)) - (Q_2(X_i', \pi(X_i')) - P^\pi Q_2(z_i)) \right] g_1(z_i) \\
&\qquad + \frac{1}{n} \sum_{i=1}^n \gamma (Q_2(X_i', \pi(X_i')) - P^\pi Q_2(z_i))(g_1(z_i) - g_2(z_i)) \\
&\le 2L \|g_1 - g_2\|_{P_n} + \gamma R \left[\|Q_1 - Q_2\|_\infty + \|P^\pi Q_1 - P^\pi Q_2\|_\infty \right] + 3\gamma L \|g_1 - g_2\|_{P_n} \\
&= (2 + 3\gamma) L \|g_1 - g_2\|_{P_n} + \gamma R \|Q_1 - Q_2\|_\infty + R \|T^\pi Q_1 - T^\pi Q_2\|_\infty,
\end{aligned} \tag{6.53}$$

where we used the boundedness assumptions, the definition of the supremum norm, the norm inequality $\frac{1}{n} \sum_{i=1}^n |g_1(z_i) - g_2(z_i)| \le \|g_1 - g_2\|_{P_n}$, and the fact that $|\gamma Q(X_i', \pi(X_i')) - \gamma P^\pi Q(z_i)| = |r(z_i) + \gamma Q(X_i', \pi(X_i')) - T^\pi Q(z_i)| \le (2 + \gamma) L \le 3L$ 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 equality.

Let $\{(g_j^s, g_j'^s)\}_{j=1}^{N_s}$ 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 $(g^s, (Q^s, T^\pi Q^s, \mathbf{1})) = (g^s, g'^s) \in \{(g_j^s, g_j'^s)\}_{j=1}^{N_s}$ such that $\|(g, g') - (g^s, g'^s)\|_\infty \le 2^{-s}R$. This implies that $\|Q^s - Q\|_\infty$

and $\|T^\pi Q^s - T^\pi Q\|_\infty$ are smaller than $2^{-s}R$ as well. Moreover, $\|g^s - g\|_{P_n} \le \|g^s - g\|_\infty \le 2^{-s}R$. By (6.53) we get

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

Choose $S = \min\{s \ge 1 : 2^{-s} \le \frac{\varepsilon}{7RL}\}$, which entails that for any $(g, g') \in \mathcal{G} \times \mathcal{G}'$, the net defined by $\{(g_j^S, g_j'^S)\}_{j=1}^{N_S}$ 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

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

We use chaining technique 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 our chaining argument by the use of the union bound and the fact that $N_s N_{s-1} \le N_s^2$ to get

$$
\begin{aligned}
P_1 &= \mathbb{P}\left\{\sup_{(g,g')\in\mathcal{G}\times\mathcal{G}'}\left|\frac{1}{n}\sum_{i=1}^n W_i(g')g(z_i)\right| \geq \delta\right\} \\
&\leq \mathbb{P}\left\{\max_{j=1,\ldots,N_S}\left|\frac{1}{n}\sum_{i=1}^n W_i({g'_j}^S)g_j^S(z_i)\right| \geq \delta-\varepsilon\right\} \\
&\leq \sum_{s=1}^S \mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n (R_i - r(z_i))(g^s(z_i)-g^{s-1}(z_i))\right| \geq \frac{\eta_s(\delta-\varepsilon)}{3}\right\} \\
&\quad + \mathbb{P}\Bigg\{\Bigg|\frac{1}{n}\sum_{i=1}^n \gamma\big[(Q^s(X'_i,\pi(X'_i)) - (P^\pi Q^s)(z_i)) - \\
&\qquad\qquad (Q^{s-1}(X'_i,\pi(X'_i)) - (P^\pi Q^{s-1})(z_i))\big]g^s(z_i)\Bigg| \geq \frac{\eta_s(\delta-\varepsilon)}{3}\Bigg\} \\
&\quad + \mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n \gamma(Q^{s-1}(X'_i,\pi(X'_i)) - (P^\pi Q^{s-1})(z_i))(g^s(z_i)-g^{s-1}(z_i))\right| \geq \frac{\eta_s(\delta-\varepsilon)}{3}\right\} \\
&\leq \sum_{s=1}^S N_s N_{s-1}\exp\left(-\frac{2(\delta-\varepsilon)^2\eta_s^2 n}{4\times 9^2(RL)^2 2^{-2s}}\right) + N_s N_{s-1}\exp\left(-\frac{2(\delta-\varepsilon)^2\eta_s^2 n}{4\times 9^2 R^4 2^{-2s}}\right) \\
&\qquad\qquad + N_s N_{s-1}\exp\left(-\frac{2(\delta-\varepsilon)^2\eta_s^2 n}{3\times 9^2(RL)^2 2^{-2s}}\right) \\
&\leq \sum_{s=1}^S \exp\left(3\log N_s - \frac{2(\delta-\varepsilon)^2\eta_s^2 n}{4\times 9^2(RL)^2 2^{-2s}}\right). \qquad (6.54)
\end{aligned}
$$

Choose

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

It can be shown that by this choice of η_s, $\sum_{s=1}^S \eta_s \leq 1$. Take C in (6.52) sufficiently large such that

$$
\sqrt{n}(\delta-\varepsilon) \geq 2\times 3^3 RL\sum_{s=1}^S 2^{-s}[\log\mathcal{N}_\infty(2^{-s}L,\mathcal{G}\times\mathcal{G}')]^{1/2} \ \vee 72\sqrt{6\log 2}\,RL. \qquad (6.55)
$$

We have $\log N_s \leq \frac{n(\delta-\varepsilon)^2\eta_s^2}{3^6(RL)^2 2^{-2s}}$, so P_1 in (6.54) can be upper bounded as follows

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

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

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

where in the last inequality we used the assumption that $\sqrt{n}(\delta-\varepsilon) \geq 72\sqrt{6\log 6}\,RL$ (cf. (6.55)).

One can show that (6.55) is satisfied if

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

so C can be chosen as $C = 36(2\sqrt{6\log 2}R \vee 1)$. □

The following lemma, which is built on Lemma 6.14, is a modulus of continuity result and will be used in the proof of Theorem 6.8. 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 $\mathcal{G}\times\mathcal{G}'$, i.e., it is a pseudo-norm.[13]

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 6.8. 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 6.15 (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 exit $0<\alpha<1$ and a finite constant A such that for all $u>0$,*

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

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

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

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 $L\geq 1$. We use $c_1, c_2, \ldots > 0$ as constants. First we start by peeling on $J(g,g')$:

$$\begin{aligned}
\delta &\triangleq \mathbb{P}\left\{\sup_{(g,g')\in\mathcal{G}\times\mathcal{G}'} \frac{|\langle W(g'), g\rangle_n|}{\|g\|_{P_n}^{1-\alpha} J^\alpha(g,g')} \geq t\right\} \\
&\leq \sum_{s=0}^{\infty} \mathbb{P}\left\{\sup_{(g,g')\in\mathcal{G}\times\mathcal{G}'} \frac{|\langle W(g'), g\rangle_n|}{\|g\|_{P_n}^{1-\alpha}} \geq \underbrace{t.2^{\alpha s}}_{\triangleq \tau_s},\ 2^s \mathbb{I}_{\{s\neq 0\}} \leq J(g,g') < 2^{s+1}\right\}. \qquad (6.56)
\end{aligned}$$

Let us denote each term in the RHS by δ_s. To upper bound δ_s, notice that by assumption

[13] The statement of this lemma is different from the corresponding result in the originally submitted dissertation in 2011.

$\|g\|_{P_n} \le L$. For each term, we peel again on $\|g\|_{P_n}$ and apply Lemma 6.14 as the following:

$$\begin{aligned}
\delta_s &\le \sum_{r\ge 0} \mathbb{P}\left\{ \sup_{(g,g')\in\mathcal{G}\times\mathcal{G}'} \frac{|\langle W(g'), g\rangle_n|}{\|g\|_{P_n}^{1-\alpha}} \ge \tau_s, 2^s \mathbb{I}_{\{s\neq 0\}} \le J(g,g') < 2^{s+1}, 2^{-(r+1)}L < \|g\|_{P_n} \le 2^{-r}L \right\} \\
&\le \sum_{r\ge 0} \mathbb{P}\left\{ \sup_{(g,g')\in\mathcal{G}\times\mathcal{G}'} |\langle W(g'), g\rangle_n| \ge \tau_s \left(2^{-(r+1)}L\right)^{1-\alpha}, J(g,g') < 2^{s+1}, \|g\|_{P_n} \le 2^{-r}L \right\} \\
&\le \sum_{r\ge 0} 2\exp\left(-\frac{n\left[\tau_s \left(2^{-(r+1)}L\right)^{1-\alpha}\right]^2}{6^7 \left(2^{-r}L\right)^2 L^2} \right) \\
&= \sum_{r\ge 0} 2\exp\left(-\frac{2^{2r\alpha} n\tau_s^2}{6^7 \times 2^{2(1-\alpha)} L^{2(1+\alpha)}} \right) = c_2 \exp\left(-\frac{c_1 n\tau_s^2}{L^{2(1+\alpha)}} \right). \qquad (6.57)
\end{aligned}$$

The last inequality holds only if the covering number condition in Lemma 6.14 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 will be satisfied for

$$t \ge \frac{CL\sqrt{A} 2^{1+\alpha}}{\sqrt{n}} \vee \frac{2^{1-\alpha} C L^{1+\alpha}}{\sqrt{n}} = c_3 \frac{L^{1+\alpha}}{\sqrt{n}}. \qquad (6.58)$$

Plug-in (6.57) in (6.56) to get that

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

Solving for δ, we have $t \le c_5 L^{1+\alpha} \sqrt{\frac{\ln(1/\delta)}{n}}$ with probability at least $1-\delta$. This alongside the condition (6.58) lead to the desired result. □

Let us turn to the proof of Theorem 6.8. 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 of Theorem 6.8. Recall that in the optimization problem, we use $w_i = (X_i, A_i, R_i, X_i')$ $(i = 1, \dots, n)$ to denote the i^{th} elements of the dataset $\mathcal{D}_n = \{(X_i, A_i, R_i, X_i')\}_{i=1}^n$. 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 $i = 1, \dots, n$, 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 $[W_1(Q) \dots W_n(Q)]^\top$. Notice that $|W_i(Q)| \le 3Q_{\max}$, and we have $\mathbb{E}[W_i(Q) \mid Q] = 0$ $(i = 1, \dots, n)$.

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

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

After expanding and rearranging, we get

$$\left\|\hat{h}_n(Q) - T^\pi Q\right\|_n^2 + \lambda_{h,n} J^2(\hat{h}_n(Q)) \leq 2 \left\langle W(Q)\,,\, \hat{h}_n(Q) - T^\pi Q \right\rangle_n + \lambda_{h,n} J^2(T^\pi Q). \quad (6.59)$$

We evoke Lemma 6.15 to upper bound $\left|\left\langle W(Q, T^\pi Q)\,,\, \hat{h}_n(Q) - T^\pi Q \right\rangle_n\right|$. 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

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

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 $(\mathcal{G} \times \mathcal{G}')_B = \{ (g, g') \in \mathcal{G} \times \mathcal{G}' \,:\, J(g, g') \leq B \}$. Lemma 6.15 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 $\mathcal{F}_B = \{ f \in \mathcal{F} \,:\, J(f) \leq B \}$, which is specified by Assumption A16.

Notice that if $J(g, g') \leq B$, each of $J(h)$, $J(Q)$, and $J(T^\pi Q)$ is also smaller than B. So we have

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

Because $J(\cdot)$ is a pseudo-norm, it holds that $J(h - T^\pi Q) \leq J(h) + J(T^\pi Q)$, so the set $\left\{ h - T^\pi Q \,:\, h, T^\pi Q \in \mathcal{F}^{|\mathcal{A}|}, J(h) + J(T^\pi Q) \leq B \right\}$ is a subset of the set $\left\{ Q \,:\, Q \in \mathcal{F}^{|\mathcal{A}|}, J(Q) \leq B \right\}$. As a result $(\mathcal{G} \times \mathcal{G}')_B$ is a subset of the product space $\left\{ Q \in \mathcal{F}^{|\mathcal{A}|} \,:\, J(Q) \leq B \right\}^3$. Therefore by the usual covering argument, we get that

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

It is easy to see that for finite $|\mathcal{A}|$, if $\log \mathcal{N}_\infty(u, \{f \in \mathcal{F} : J(f) \leq B\}) \leq C(\frac{B}{u})^{2\alpha}$, then $\log \mathcal{N}_\infty(u, \{(f_1, \ldots, f_{|\mathcal{A}|}) \in \mathcal{F}^{|\mathcal{A}|} : J\left((f_1, \ldots, f_{|\mathcal{A}|})\right) \leq B\}) \leq C_1(\frac{B}{u})^{2\alpha}$ (we benefit from the condition $J(Q(\cdot, a)) \leq J(Q)$ in Assumption A15; the proof is similar to that of Lemma 6.E). Here the constant C_1 depends on $|\mathcal{A}|$. This along the previous inequality show that for some constant $A > 0$, we have

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

We are ready to apply Lemma 6.15 to upper bound the inner product term in (6.59). 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 constant $c > 0$ such that with probability at least $1 - \delta$, it holds that

$$L_n^2 + \lambda_{h,n} J^2(\hat{h}_n) \leq 2cL^{1+\alpha} L_n^{1-\alpha} \left(J(\hat{h}_n) + J(Q) + J(T^\pi Q) \right)^\alpha t_0 + \lambda_{h,n} J^2(T^\pi Q). \quad (6.60)$$

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

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

$$L_n^2 + \lambda_{h,n} J^2(\hat{h}_n) \leq 4cL^{1+\alpha} L_n^{1-\alpha} \left(J(\hat{h}_n) + J(Q) + J(T^\pi Q) \right)^\alpha t_0. \quad (6.61)$$

Again, two cases might happen:

Case 1.a. $J(\hat{h}_n) > J(Q) + J(T^\pi Q)$: From (6.61) we have $L_n^2 \leq 2^{2+\alpha} cL^{1+\alpha} L_n^{1-\alpha} J^\alpha(\hat{h}_n) t_0$. 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 (6.61) we also have $\lambda_{h,n} J^2(\hat{h}_n) \leq 2^{2+\alpha} cL^{1+\alpha} L_n^{1-\alpha} J^\alpha(\hat{h}_n) t_0$. Plugging-in the recently obtained upper bound on L_n and solving for $J(\hat{h}_n)$, we get that

$$J(\hat{h}_n) \leq \frac{2^{2+\alpha} cL^{1+\alpha} t_0}{\lambda_{h,n}^{\frac{1+\alpha}{2}}}. \quad (6.62)$$

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

$$L_n \leq \frac{2^{2+\alpha} cL^{1+\alpha} t_0}{\lambda_{h,n}^{\frac{\alpha}{2}}}. \quad (6.63)$$

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 (6.61) we have $L_n^2 \leq 2^{2+\alpha} cL^{1+\alpha} L_n^{1-\alpha} (J(Q) + J(T^\pi Q))^\alpha t_0$. Solving for L_n, we obtain

$$L_n \leq 2^{\frac{2+\alpha}{1+\alpha}} c^{\frac{1}{1+\alpha}} L \left(J(Q) + J(T^\pi Q) \right)^{\frac{\alpha}{1+\alpha}} t_0^{\frac{1}{1+\alpha}}. \quad (6.64)$$

Case 2. $2cL^{1+\alpha} L_n^{1-\alpha} (J(\hat{h}_n) + J(Q) + J(T^\pi Q))^\alpha t_0 < \lambda_{h,n} J^2(T^\pi Q)$. 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

$$L_n \leq \sqrt{2\lambda_{h,n}} J(T^\pi Q), \quad (6.65)$$

$$J(\hat{h}_n) \leq \sqrt{2} J(T^\pi Q). \quad (6.66)$$

By (6.63), (6.64), and (6.65) for L_n and (6.62), (6.66), 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 $1 - \delta$, it holds that

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

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

□

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 6.15, 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, the first term in the RHS of (6.60) 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. 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 $g' \in \mathcal{G}'$.

6.E Proof of Lemma 6.12 (Covering number of G_{σ_1,σ_2})

Here we prove Lemma 6.12, 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 of Lemma 6.12. For any $g_{Q_1,h_1}, g_{Q_2,h_2} \in G_{\sigma_1,\sigma_2}$ and $z_i = (x_i, a_i)$, we have

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

Assumption A15 implies that $Q_{1,j}, Q_{2,j} \in \mathcal{F}_{\sigma_1}$ and $h_{1,j}, h_{2,j} \in \mathcal{F}_{\sigma_2}$ for all $j = 1, \cdots, |\mathcal{A}|$. Therefore, an u-cover 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 $\|\cdot\|_{x_{1:n}}$ defines an $8Q_{\max}\sqrt{|\mathcal{A}|}$ u-cover on G_{σ_1,σ_2} w.r.t. $\|\cdot\|_{z_{1:n}}$. Thus,

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

Assumption A16 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

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

□

6.F Why Two Regularizers?

We discuss why using regularizers in both optimization problems (6.13) and (6.14) of REG-LSTD is necessary for large function spaces such as Sobolev spaces and 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 (unmodified) BRM, which is biased, or the solution can be arbitrary bad.[14]

[14] This section was not present in the originally submitted dissertation in September 2011. It was added later in the Fall of 2014.

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 function w.r.t. the supremum norm. This is satisfied by many large function spaces such as RKHSs with universal kernels (Definition 4.52 of Steinwart and Christmann 2008). We consider what would happen if instead of the current formulation of REG-LSTD, which is

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

we only used a regularizer either in the first or second optimization. We study each case separately.

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

$$\hat{h}_n(\cdot;Q) = \operatorname*{argmin}_{h\in\mathcal{F}^{|\mathcal{A}|}} \left\| h - \hat{T}^\pi Q \right\|_{\mathcal{D}_n}^2,$$
$$\hat{Q} = \operatorname*{argmin}_{Q\in\mathcal{F}^{|\mathcal{A}|}} \left[\left\| Q - \hat{h}_n(\cdot;Q) \right\|_{\mathcal{D}_n}^2 + \lambda_{Q,n} J^2(Q) \right]. \tag{6.67}$$

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 $i = 1, \ldots, n$.[15] Such a function is indeed the minimizer of the loss $\|Q - \hat{h}_n(\cdot;Q)\|_{\mathcal{D}_n}^2$. The second optimization problem (6.67) becomes

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

This is indeed the regularized version of the original (i.e., unmodified) formulation of the BRM algorithm. As discussed in Section 6.2.1, the unmodified BRM algorithm is biased when the MDP is not deterministic. Adding a regularizer does not solve the biasedness problem of the 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 $\|h - \hat{T}^\pi Q\|_{\mathcal{D}_n}^2$, but we do not regularize $\|Q - \hat{h}_n(\cdot;Q)\|_{\mathcal{D}_n}^2$, i.e.,

$$\hat{h}_n(\cdot;Q) = \operatorname*{argmin}_{h\in\mathcal{F}^{|\mathcal{A}|}} \left[\left\| h - \hat{T}^\pi Q \right\|_{\mathcal{D}_n}^2 + \lambda_{h,n} J^2(h) \right],$$
$$\hat{Q} = \operatorname*{argmin}_{Q\in\mathcal{F}^{|\mathcal{A}|}} \left\| Q - \hat{h}_n(\cdot;Q) \right\|_{\mathcal{D}_n}^2. \tag{6.68}$$

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 6.7 and Theorem 6.8). So we can expect $\hat{h}_n(\cdot;Q)$ to get closer to $T^\pi Q$ as the sample size grows.

[15] To be more precise: First for an $\varepsilon > 0$, we construct a continuous function $\bar{h}_\varepsilon(z) = \sum_{Z_i \in \{(X_i,A_i)\}_{i=1}^n} \max\left\{1 - \frac{\|z - Z_i\|}{\varepsilon}, 0\right\} (\hat{T}^\pi Q)(Z_i)$. This construction is similar to Theorem 2 of Nadler et al. [2009]. 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 $\hat{h}_n(\cdot;Q) = h_\varepsilon$.

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 is we precisely know $T^\pi Q$ at all data points. Substituting this $\hat{h}_n((x,a);Q)$ in the second optimization problem (6.67), we get that we are indeed solving the following optimization problem:

$$\hat{Q} = \operatorname*{argmin}_{Q \in \mathcal{F}^{|\mathcal{A}|}} \|Q - T^\pi Q\|_{\mathcal{D}_n}^2 \tag{6.69}$$

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

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

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 $\bar{Q}_\varepsilon$. For small enough ε, the function Q_ε is a minimizer of (6.69), i.e., the value of $\|Q_\varepsilon - T^\pi Q_\varepsilon\|_{\mathcal{D}_n}^2$ 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.

6.G Convolutional MDPs and Assumption A19

In this appendix, we show that Assumption A19 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 the function of chosen action only, i.e., $X' - X \sim W(\pi(X))$.

Proposition 6.16. *Suppose that $\mathcal{X} = [-\pi, \pi]$ is the unit circle and $\mathcal{F}$ is the Sobolev space $\mathcal{W}^{k,2}(\mathcal{X})$ 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 A19 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 $\|V\|_{\mathcal{W}^{k,2}}^2 = \sum_{n=-\infty}^{\infty} \left(1 + |n|^2\right)^k |\tilde{V}(n)|^2$. Thus,

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

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. □

Chapter 7

Model Selection in Reinforcement Learning

7.1 Introduction

One of the most important open questions in reinforcement learning is how to optimally choose the function approximation architecture and its parameters for a given problem. For instance, a designer must decide about the number, the location, and the shape of basis functions of a radial basis function network; the number of layers and neurons in a neural network; or the number of tilings and their resolutions in a tile coding (cf. Chapter 8 of the book of Sutton and Barto [1998]). Other examples are the regularization coefficient and/or kernel parameters of a regularized kernel-based reinforcement learning algorithm (Chapters 5 and 6 of this work; Engel et al. 2005; Jung and Polani 2006; Loth et al. 2007; Taylor and Parr 2009; Kolter and Ng 2009), or other parameters that directly or indirectly determine the function approximation architecture (e.g., the parameters of the evolutionary algorithm NEAT in Whiteson and Stone 2006). At an even higher level, one has to decide about which of these function approximation methods should be used. All these can be represented as a choice of parameters, if the word parameter is understood in a sufficiently general sense.[1]

It is widely recognized that the best choice is problem dependent. Hence, it makes sense to choose the parameters data-dependently with the ultimate goal of picking them such that the resulting performance is almost as good as if the algorithm's best, but unknown, problem-dependent parameter setting was used.

In this chapter we study the problem of automatic parameters tuning in the offline sampling scenario (Section 2.2) for MDPs. This problem is difficult because in the offline sampling scenario there is no direct way to evaluate the performance of a given policy. We investigate parameter tuning when we want to find a good approximation to the fixed point of the Bellman optimality operator.

To make the goal of systematic parameter tuning clear, consider the following setting: Assume that we are given a learning algorithm A that takes the data $\mathcal{D}_n$ and a parametrized function space $\mathcal{F}^{|\mathcal{A}|}(p)$ and then proposes an action-value function $Q_n = \mathsf{A}(\mathcal{D}_n, \mathcal{F}^{|\mathcal{A}|}(p))$ which is an element of $\mathcal{F}^{|\mathcal{A}|}(p)$. The task of A, ideally, is to come up with a function $Q_n \in \mathcal{F}^{|\mathcal{A}|}(p)$ whose Bellman error is close to that of the best possible choice from $\mathcal{F}^{|\mathcal{A}|}(p)$. For example, the algorithm A might be RFQI (Chapter 5) or REG-LSPI/BRM (Chapter 6). And the parameter p might be the regularization coefficients and the parameter describing the kernel function.

Suppose p^* is the unknown parameter for the algorithm A on a given problem that achieves the smallest Bellman error. The goal of this work is to design a parameter-tuning

[1] This chapter is the result of the collaboration of the author with Csaba Szepesvári.

algorithm that, given the data, chooses a parameter $\hat{p}$ such that the resulting performance is (almost) as good as the performance of the algorithm A running on the data with $\mathcal{F}^{|\mathcal{A}|}(p^*)$. If a parameter-tuning algorithm achieves this goal, we call it *adaptive*.

7.1.1 Contributions

In supervised learning, a classical method to achieve adaptivity is *complexity regularization* [Barron, 1991; Bartlett et al., 2002; Wegkamp, 2003; Lugosi and Wegkamp, 2004]. A straightforward adoption of complexity regularization to our problem suggests the following procedure:

Assume that the possible parameter settings are enumerated in a list $p_1, p_2, \ldots$. For $k = 1, 2, \ldots$, run algorithm A using the function space $\mathcal{F}^{|\mathcal{A}|}(p_k)$ to obtain an action-value function candidate $Q_k = \mathsf{A}(\mathcal{D}_n, \mathcal{F}^{|\mathcal{A}|}(p_k))$. Next, estimate the Bellman error of Q_k, e.g., using a hold-out data with n observations. Let the resulting estimate be $\mathrm{BE}_n(Q_k)$. Then choose

$$\hat{k} = \operatorname*{argmin}_{k \geq 1} \left[C_1 \, \mathrm{BE}_n(Q_k) + C_2 \, \frac{\log k}{n} \right],$$

where $C_1 \geq 1$ and $C_2 > 0$ are well-chosen constants. Generic model selection results can then be used to show that this procedure is indeed adaptive, provided that $\mathrm{BE}_n(Q_k)$ is an unbiased estimate of the Bellman error of Q_k (Theorem 7.1).

Unfortunately, we know of no way to derive an unbiased estimate of the Bellman error of Q_k based on a finite amount of data. Therefore the above procedure, which is standard in supervised learning setting, is not adequate for reinforcement learning problems.

The main algorithmic contribution of the work is a method called BErMin, which is similar to the above mentioned procedure, but can in fact be implemented and still achieves adaptivity. This method overcomes the difficulty of not being able to measure the Bellman error directly. We discuss BErMin in Section 7.3, and provide an intuitive explanation of why the algorithm works (Section 7.3.1). After the pseudo-code of the algorithm is presented in Section 7.3.2, we give an example of how BErMin may be used in conjunction with a standard reinforcement learning algorithm, such as LSPI (Lagoudakis and Parr [2003]), in order to make an almost parameter-free meta-algorithm (Section 7.3.3).

The main theoretical contribution of this work is Theorem 7.2 that shows that BErMin has an oracle-like property (Section 7.4.2), in the sense that it selects the model with the minimum Bellman error up to a multiplicative constant and some additional terms that converge to zero. This indeed implies that the procedure is adaptive in a sense that will be precisely defined (Theorem 7.3 in Section 7.4.3).

In addition to these main contributions, we provide some auxiliary results that might have applications in more general than reinforcement learning context. In particular, Theorem 7.1 is an umbrella result for complexity-regularization-based model selection, and its application leads to Theorem 7.2. This theorem is a generalized form of Theorem 3 of Bartlett et al. 2002 with some differences that we discuss in Section 7.4.1. Later on in the appendix, we provide a noncentral tail inequalities for Hidden Markov Processes that helps us to obtain faster rates by controlling the variance of a random variable (Lemma 7.4 and Lemma 7.7 in Section 7.C). Finally, in Section 7.D we provide a procedure to estimate the excess error of a regression problem. Interesting on its own, this result will be used in BErMin.

7.2 Problem Definition

Suppose that we are given a list of action-value functions $Q_1, Q_2, \ldots, Q_P$ (with the possibility of $P > n$, or even $P = \infty$) and a dataset $\mathcal{D}_n$, the latter satisfying the standard offline sampling assumption. Our goal is to devise a procedure that selects the action-value function

amongst $\{Q_1, \ldots, Q_P\}$ that has the smallest (integrated, squared) Bellman (optimality) error. Thus, the ideal procedure would return $Q_{\hat{k}}$, where

$$\hat{k} = \operatorname*{argmin}_{1 \le k \le P} \|Q_k - T^* Q_k\|_\nu^2 .$$

The idea of using the Bellman error as a criterion of optimization is not new. The algorithms implementing generalized policy iteration can be viewed as working towards minimizing it, e.g., Lagoudakis and Parr [2003]; Antos et al. [2008b]. There are also some basis generation/adaptation methods that use the Bellman error to guide their search, e.g., Menache et al. [2005]; Keller et al. [2006]; Parr et al. [2007]. For a justification of minimizing the Bellman error see the discussion in the paper by Antos et al. [2008b] following their Theorem 4, or Lemma 7 of Antos et al. [2007].

Unfortunately, the Bellman error is not easy to work with. This is because neither T^* nor T^π is available in the learning setting. Moreover, even though $\hat{T}^*$ ($\hat{T}^\pi$) provides an unbiased estimate to T^* (respectively, T^π) in the sense of Proposition 2.1, these operators cannot be used in a simple manner to estimate the Bellman error. One might think that given any fixed function Q, the mean-squared empirical Bellman residual, $\|Q - \hat{T}^* Q\|_n^2$, is a reasonable estimate to the Bellman error. However, it follows from a standard bias-variance decomposition that

$$\mathbb{E}\left[\|Q - \hat{T}^* Q\|_n^2\right] = \|Q - T^* Q\|_\nu^2 + \mathbb{E}\left[\|\hat{T}^* Q - T^* Q\|_n^2\right] \neq \|Q - T^* Q\|_\nu^2 ,$$

which shows that $\|Q - \hat{T}^* Q\|_n^2$ is a biased estimate. In fact, from the above decomposition, we see that selecting the policies based on the mean-squared empirical Bellman residual leads to favoring policies whose underlying variance-like term $\mathbb{E}\left[\|T^* Q - \hat{T}^* Q\|_n^2\right]$ is small, as noted previously by, e.g., Menache et al. [2005] or Antos et al. [2008b].

The main contribution of this work is a procedure, BErMin, and its analysis that shows that BErMin finds a candidate whose Bellman error is not much larger than that of the best candidate.

Remark 7.1. In the analysis below, for the sake of simplicity, we assume that $Q_1, \ldots, Q_P$ are fixed deterministic functions. In practice, these functions would be estimated based on some data, in which case, they would become random (data-dependent) functions. Our results, however, still continue to hold provided that the sample $\mathcal{D}_n$ used to evaluate the candidates is independent of $Q_1, \ldots, Q_P$. In particular, in this case the results can be stated and proven on the probability space obtained by conditioning on the data that generated $Q_1, \ldots, Q_P$ (the proofs would work word-by-word with no further changes). The study of the case when the same data is used to generate $Q_1, \ldots, Q_P$ is left for future work. One possible starting point for such a study could be the work by Antos et al. [2008b], who have analyzed the theoretical properties of approximate policy iteration when the same data is used in all iterations, with the main message of their result being that the correlations arising from reusing the same data are not necessarily catastrophic.

7.3 Model Selection Algorithm for Bellman Error Minimization (BErMin)

The purpose of this section is to introduce BErMin, a complexity regularization-based model selection algorithm for the problem of finding the Bellman error minimizer among the action-value function candidates $\{Q_k\}_{k=1}^P$. The setup is as described in Section 7.2. We start by describing the main idea behind the algorithm in Section 7.3.1, while the algorithm itself is presented in Section 7.3.2. Finally in Section 7.3.3, we show an example of how BErMin may be used to devise an almost parameter-free reinforcement learning algorithm by modifying the conventional LSPI algorithm.

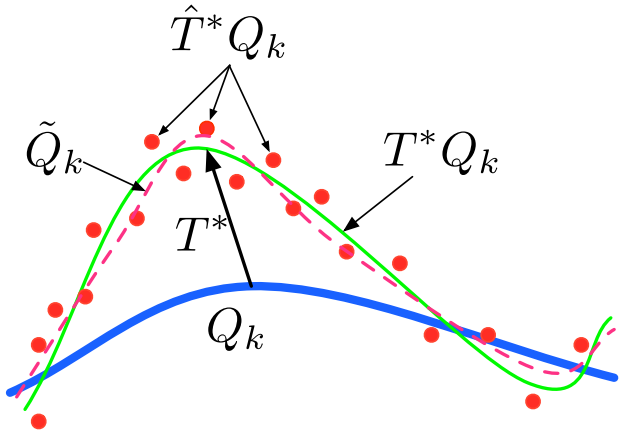

Figure 7.1: When the problem is to estimate the difference between T^*Q_k (solid green line) and Q_k (bold, solid blue line) and the function T^*Q_k is unknown, one may use samples from $\hat{T}^*Q_k$ (red dots) and solve a regression problem to get $\tilde{Q}_k$ (dashed red line). This estimate can be used in place of T^*Q_k to construct an estimate of $T^*Q_k - Q_k$.

7.3.1 The Idea Behind the Algorithm

The basic idea behind our approach is that while the Bellman operator T^* itself is not accessible, one still may approximately learn T^*Q and use it to estimate the Bellman error. Thanks to the definition of the empirical Bellman operator $\hat{T}^*$ (Definition 2.8), the regression function underlying

$$\mathcal{D}_{n,k} = \left\{ \Big((X_1, A_1), (\hat{T}^*Q_k)(X_1, A_1) \Big), \dots, \Big((X_n, A_n), (\hat{T}^*Q_k)(X_n, A_n) \Big) \right\} \tag{7.1}$$

is T^*Q_k (cf. Proposition 2.1). Thus, we can feed $\mathcal{D}_{n,k}$ to a regression procedure which, ideally, returns a "good" approximation to T^*Q_k. As the regression method one can use any of the large number of state-of-the-art techniques, such as the regularized least-squares regression algorithm of Chapter 4 (cf., the books by Hastie et al. 2001; Györfi et al. 2002; Wasserman 2007; Rasmussen and Williams 2006; Bishop 2006). Although the discussion of the relative merits of the available methods is beyond the scope of this chapter, we will shortly be more specific about the desired properties of the method.

Let the action-value function returned by the chosen regression algorithm be denoted by $\tilde{Q}_k$. If $\tilde{Q}_k$ is close to T^*Q_k, then by calculating $\|Q_k - \tilde{Q}_k\|_n^2 \approx \|Q_k - \tilde{Q}_k\|_\nu^2 \approx \|Q_k - T^*Q_k\|_\nu^2$ one can select the action-value function with the smallest Bellman error based on computing

$$\operatorname*{argmin}_{1 \le k \le P} \|Q_k - \tilde{Q}_k\|_n^2 .$$

Figure 7.1 depicts function $\tilde{Q}_k$ and its relation to Q_k and T^*Q_k.

The problem with this procedure is that it might be overly optimistic and thus it may result in an uncontrolled error. To see why, imagine that for some index k_0 whose associated Bellman error $\|Q_{k_0} - T^*Q_{k_0}\|_\nu^2$ is "large", the regression procedure returns an estimate such that $\|Q_{k_0} - \tilde{Q}_{k_0}\|_\nu^2 \ll \|Q_{k_0} - T^*Q_{k_0}\|_\nu^2$ (for example, because the regression procedure might be biased towards action-values close to zero, Q_{k_0} might be close to zero, while $T^*Q_{k_0}$ might be far from zero, cf. also Figure 7.2). As a result, the above procedure will likely select k_0, and thus might miss some other index with a lower Bellman error. To avoid this problem, we must guard the procedure against the underestimation of the Bellman error.

This situation is illustrated in Figure 7.2. The function T^*Q_k is once approximated by $\tilde{Q}_k^{(1)}$ and the other time by $\tilde{Q}_k^{(2)}$. For $\tilde{Q}_k^{(2)}$, the value of $\|Q_k - \tilde{Q}_k^{(2)}\|_\nu$ is small, even though

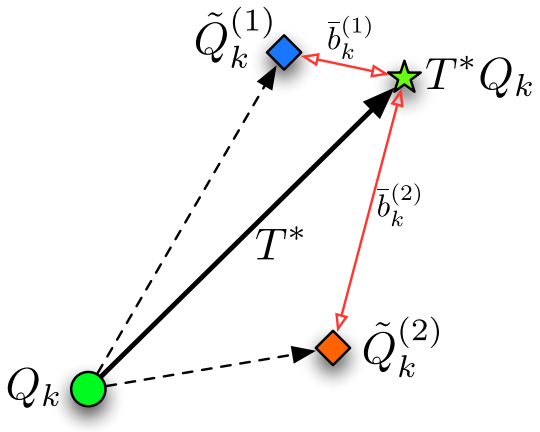

Figure 7.2: Consider the problem of estimating the Bellman error $\|Q_k - T^*Q_k\|_\nu^2$. If T^*Q_k is replaced by a surrogate $\tilde{Q}_k^{(1)}$, $\|Q_k - \tilde{Q}_k^{(1)}\|_\nu^2$ gives a relatively good estimate of this quantity because $\tilde{Q}_k^{(1)}$ is close to T^*Q_k. However, when $\tilde{Q}_k^{(2)}$ replaces T^*Q_k, the resulting estimate of the Bellman error becomes poor and $\|Q_k - \tilde{Q}_k^{(2)}\|_\nu^2$ would be an *underestimate* of the true Bellman error. This might lead to the unjust selection of the candidate Q_k. One way to protect oneself against such mistakes is to take into account how well the surrogate $\tilde{Q}_k$ approximates T^*Q.

it is a bad estimate of $\|Q_k - T^*Q_k\|_\nu$. On the other hand, $\|Q_k - \tilde{Q}_k^{(1)}\|_\nu$ is a larger number but provides a better estimate of the true Bellman error.

BErMin achieves this by correcting $\|Q_k - \tilde{Q}_k\|_\nu^2$ with $\|T^*Q_k - \tilde{Q}_k\|_\nu^2$. Since

$$\|Q_k - T^*Q_k\|_\nu^2 \le 2\left[\|Q_k - \tilde{Q}_k\|_\nu^2 + \|T^*Q_k - \tilde{Q}_k\|_\nu^2\right],$$

the correction indeed prevents the choice of an overly optimistic estimate (the sum in the brackets cannot be less than half of the estimated quantity). The first term of the right-hand side can be estimated by $\|Q_k - \tilde{Q}_k\|_n^2$. We further assume that we are provided with a (tight) high-probability upper bound, $\bar{b}_k$, on $\|T^*Q_k - \tilde{Q}_k\|_\nu^2$, i.e., $\|T^*Q_k - \tilde{Q}_k\|_\nu^2 \le \bar{b}_k$ with high probability. We propose to select the action-value function corresponding to the minimum of $\|Q_k - \tilde{Q}_k\|_n^2 + \bar{b}_k$. If $\bar{b}_k$ is a sufficiently tight bound, we expect that using $\bar{b}_k$ in place of $\|T^*Q_k - \tilde{Q}_k\|_\nu^2$ will not introduce any significant further bias. Going back to the example in Figure 7.2, the value of $\bar{b}_k$ corresponding to $\tilde{Q}_k^{(2)}$ is large, so it can compensate the underestimation caused by $\|Q_k - \tilde{Q}_k^{(2)}\|_\nu$.

We want to take care of one more detail. We would like our procedure to handle situations where the number of candidate action-value functions, P, is very large, or even potentially infinite. The latter situation arises when one transforms the algorithm into an anytime method, whose computation budget may or may not be limited, which keeps generating candidates if given more time. As a consequence of this, we add another penalty term that prevents optimistic selection bias and we will let $P = \infty$. If P is finite and small compared to n, this penalty term can safely be ignored.

7.3.2 BErMin Algorithm

BErMin, shown as Algorithm 3, implements the ideas described in the previous section. A graphical illustration of the procedure is given on Figure 7.3.

The algorithm's inputs are the candidate action-value functions, the dataset $\mathcal{D}_{(m,n)}$, a regression procedure Regress, a desired error probability δ, and three constants: $0 < a < 1$,

Algorithm 3 BERMIN$(\{Q_k\}_{k=1,2,\dots}, \mathcal{D}_{(m,n)}, \text{REGRESS}(\cdot), \delta, a, B, \tau)$

1: Split $\mathcal{D}_{(m,n)}$ into two disjoint parts: $\mathcal{D}_{(m,n)} = \mathcal{D}'_m \cup \mathcal{D}''_n$.
2: Choose (C_k) such that $S = \sum_{k\geq 1} \exp(-\frac{(1-a)^2 a\, n}{16B^2\tau(1+a)} C_k) < \infty$.
3: Choose (δ'_k) such that $\sum_{k\geq 1} \delta'_k = \delta/2$.
4: **for** $k = 1, 2, \dots$ **do**
5: $\quad (\tilde{Q}_k, \bar{b}_k) \leftarrow \text{REGRESS}(\mathcal{D}'_{m,k}, \delta'_k)$
6: $\quad e_k \leftarrow \frac{1}{|\mathcal{D}''_n|} \sum_{(X,A)\in\mathcal{D}''_n} (Q_k(X,A) - \tilde{Q}_k(X,A))^2$
7: $\quad \mathcal{R}_k^{\text{RL}} \leftarrow \frac{1}{(1-a)^2}\, e_k + \bar{b}_k$
8: **end for**
9: $\hat{k} \leftarrow \text{argmin}_{k\geq 1} \left[\mathcal{R}_k^{\text{RL}} + C_k\right]$
10: **return** $\hat{k}$

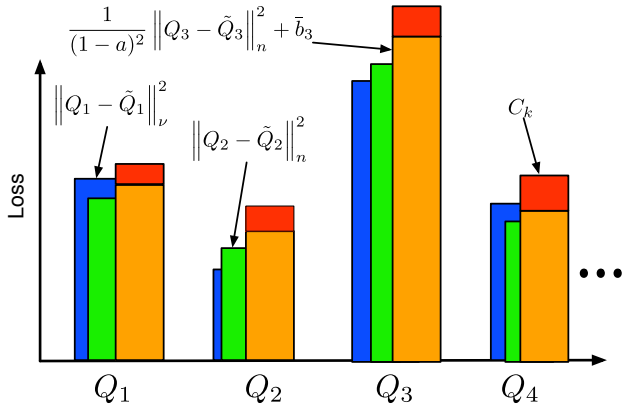

Figure 7.3: A graphical illustration of the BERMIN algorithm. The error $\|Q_k - \tilde{Q}_k\|_\nu$ (blue, leftmost bar) is estimated by $\left\|Q_k - \tilde{Q}_k\right\|_n^2$ (green, second bar from left), this is topped by $\bar{b}_k$, an upper bound on $\|\tilde{Q}_k - T^*Q_k\|_\nu^2$. This is followed by inflating this result by a factor of $\frac{1}{(1-a)^2}$ (brown, third bar, dark segment). Finally, the algorithm adds a complexity regularization term C_k (e.g., $C_k = \frac{32B^2\tau(1+a)}{a(1-a)^2 n}\ln(k)$) (red, third bar), and the minimum of all these values will be selected. In this figure, BERMIN would select the function Q_2.

B, and τ. Here a is a tuning parameter, the constant B is the bound on all functions involved (that is Q_k, $\tilde{Q}_k$, T^*Q_k, and $\bar{b}_k$), and τ is the forgetting time of the Markov chain (cf. Definition 7.1 in Appendix 7.B). The effect of these values on the quality of the solution is quantified in Theorem 7.2.

The algorithm initializes its data structures in three steps. In the first line the dataset is split into two disjoint parts, the first having m points, the second having n points. In Line 2, the values of the constants (C_k) are chosen such that they satisfy a Kraft-McMillan-like inequality

$$\sum_{k\geq 1} \exp\left(-\frac{(1-a)^2 a\, n}{16B^2\tau(1+a)} C_k\right) < \infty.$$

One feasible choice is $C_k = \frac{32B^2\tau(1+a)}{(1-a)^2 an}\ln(k)$, but any other choice is possible as long as it satisfies the required condition. The choice of these values should reflect one's prior beliefs about the suitability of the candidate functions. The default choice above (which increases with k) reflects the prior belief that functions with higher indices are less suitable. Such a choice can be justified, e.g., if Q_k is expected to become more susceptible to overfitting as

the value of k increases. When one has a finite number of models (i.e., $P < \infty$) and no good prior knowledge about the suitability of Q_k, one can use $C_k \equiv const.$ In Line 3, we choose the confidence parameters (δ'_k) such that their sum is $\delta/2$. One possible choice is $\delta'_k = \frac{3}{\pi^2}\frac{\delta}{k^2}$ (when P is finite, one can simply use $\delta'_K = \delta_k/P$). For consistency, it might be a good idea to make δ_k and C_k behave "similarly" as a function k.

In Line 5 the regression procedure REGRESS is called with the dataset $\mathcal{D}'_{m,k}$ derived from $\mathcal{D}'_m$ using (7.1) (i.e., $\mathcal{D}'_{m,k}$ depends on Q_k) and δ'_k as the confidence parameter. The requirement on REGRESS is that it returns $\tilde{Q}_k$, an estimate of T^*Q_k, and $\bar{b}_k$, a high-probability upper bound on the *excess risk* $\|\tilde{Q}_k - T^*Q_k\|_\nu^2$. The upper bound on the excess risk is required to hold with probability at least $1-\delta'_k$ (cf. Assumption A20). One possible approach to estimate the excess risk is proposed in Section 7.D.

In Line 6, the dataset $\mathcal{D}''_n$ is used to empirically estimate $\|Q_k - \tilde{Q}_k\|_\nu^2$, i.e., the blue bars in Figure 7.3 are estimated by the green bars. The error of this is expected to be well controlled (and "small"). In the next line the two error estimates are combined to yield $\mathcal{R}_k^{\mathrm{RL}}$ (brown bars in Figure 7.3). In Line 9 this estimate is further biased upwards (red portion of bars in the graph) by the amount of C_k and then the minimizer of $\mathcal{R}_k^{\mathrm{RL}} + C_k$ is selected, where $k = 1, 2, \ldots,$ giving rise to the value returned by the procedure.

Remark 7.2 (Computational Complexity). The complexity of BERMIN is expected to be dominated by the cost of running REGRESS. Let us assume that BERMIN selects the candidate returned amongst P candidates. If the computational complexity of REGRESS is $O(\mathbf{r}(m))$, the computational complexity of BERMIN becomes $O((n+\mathbf{r}(m))P)$. Thus, knowing the amount of time available, one could come up with an estimate of how many models can be evaluated. However, we think that a better approach is to run the algorithm in an anytime fashion until the computational budget is exhausted. Although BERMIN is not expected to be cheap, overall it might still be cheaper than an ad-hoc tuning method with a human in the loop, though admittedly, this would be hard to measure in practice.

Remark 7.3 (Candidate Models: An Example). An important question is what candidate functions one should feed to BERMIN and how these are found. In general, this will depend on what *a priori* information one has about the unknown MDP. Even though this is not the focus of this work, we give an example when we assume *a priori* that the optimal action-value function belongs to a Sobolev space (Definition B.3 in Appendix B.1), but the identity of the Sobolev space to which the function belongs is unknown.

For a pair $(k, J) \in \mathbb{N} \times \mathbb{R}_+$, define

$$\mathcal{F}(k, J) = \{\, f \in \mathbb{W}^k(\mathbb{R}^d \times \mathcal{A}) \,:\, \|f\|_{\mathbb{W}^k(\mathbb{R}^d \times \mathcal{A})} \le J \,\}\,.$$

Note that $\cup_{k\in\mathbb{N}, J\in\mathbb{R}_+} \mathcal{F}(k, J)$ is a huge space. For regression problems, it is known that the minimax optimal rate of estimating functions belonging to $\mathcal{F}(k, J)$ is $O(J^{2d/(2k+d)} m^{-2k/(2k+d)})$ [Györfi et al., 2002]. Here, m is the number of samples used in the learning procedure and although we use the same letter to denote the number of samples as in $\mathcal{D}'_m$, this should be considered as a coincidence.

Assume now that the true action-value function belongs to $\mathcal{F}(k^*, J^*)$ for some unknown $(k^*, J^*) \in \mathbb{N} \times \mathbb{R}_+$. Define the set of candidate function spaces as $(\mathcal{F}(k, J))_{(k,J)\in\mathcal{P}_m}$, where

$$\mathcal{P}_m = \left\{ (k, J) \in \mathbb{N} \times \mathbb{N} \,:\, \left\lceil \tfrac{d}{2} \right\rceil \le k \le m,\ J \in \{2^0, 2^1 \ldots, 2^{\lceil \log_2 m \rceil}\} \right\}.$$

This set defines a grid on both the smoothness order k and the size of the smoothness term J. As we see shortly, the resolution of this grid is set such that $\mathcal{F}(k^*, J^*)$ is contained within a member of $(\mathcal{F}(k, J))_{(k,J)\in\mathcal{P}_m}$ that is not much larger than $\mathcal{F}(k^*, J^*)$ itself.

Suppose that we have a learning algorithm A that can be configured to seek the estimate of the action-value function in $\mathcal{F}(k, J)$ and has the convergence rate of $O(J^{2d/(2k+d)} m^{-2k/(2k+d)})$, provided that the true optimal action-value function indeed belongs to $\mathcal{F}(k, J)$ (for instance, the algorithms of Chapters 5 and 6 are guaranteed to have such an optimal dependence on the number of samples). Construct $Q_{(k,J)} = \mathsf{A}(\mathcal{D}_m, \mathcal{F}(k, J))$ for all $(k, J) \in \mathcal{P}_m$. Note that

for m large enough there is a pair (k', J') in $\mathcal{P}_m$, close to (k^*, J^*), such that $\mathcal{F}(k^*, J^*)$ is contained within $\mathcal{F}(k', J')$. In particular if $m \geq \max\{k^*, J^*\}$, then there exists $(k', J') \in \mathcal{P}_m$ such that $k' = k^*$, $J' \leq 2J^*$, and $\mathcal{F}(k^*, J^*) \subset \mathcal{F}(k', J')$. The convergence rate of the estimator based on $(k', J')[= (k^*, J')]$ is $O(J'^{2d/(2k^*+d)} m^{-2k^*/(2k^*+d)})$, which is to be compared with the optimal rate, $O(J^{*2d/(2k^*+d)} m^{-2k^*/(2k^*+d)})$. We see that asymptotically, the rate associated with the model (k', J') is within at most a factor of 2 of the optimal rate. Thus, even when the set of models is restricted to a set with less than $m(\log_2(m) + 1)$ elements, by selecting an appropriate model amongst them, one can match the asymptotic rate of the true model, up to a constant factor. Thus, if we can prove that the model selected by BERMIN is almost as good as (k', J') in terms of its Bellman error, we get that BERMIN also comes within a constant factor of the Bellman error of the best model. This is the subject of Theorem 7.2, which will be stated in the next section.

7.3.3 Adaptive Linear LSPI

The choice of function approximation is crucial for the success of many RL algorithms. BERMIN, as a model selection algorithm, may be used to automate this procedure. The result would be an almost "parameter-free" meta-algorithm that data-dependently choose the right function approximator for a given problem. In this section we suggest a potential way that LSPI (Lagoudakis and Parr 2003) with linear function approximator may be modified in order to automatically select the right function approximator. A similar approach may be used for other algorithms too.

The crucial point to notice is that the right choice of function approximator depends on many factors including *(i)* the MDP itself, *(ii)* the number of available samples, and *(iii)* the iteration number of LSPI. The dependence of MDP is evident: different problems have different optimal value functions and therefore require different function approximations. Moreover, when we do not have many data samples, it is a bad idea to try to use a complicated function approximator, and vice versa. The right function approximator also depends on the iteration number of LSPI since the change of the policy leads to the change of its value function and the corresponding function space. These suggest that, computational burden aside, any change in the number of samples or the iteration of LSPI demands a new round of model selection. The modified LSPI algorithm is shown in Algorithm 4.

Algorithm 4 is very similar to the original LSPI algorithm of Lagoudakis and Parr 2003. The major difference is that it gets a set of basis functions $\{\Phi_k\}_{k=1}^P$ in which P is the number of potential function approximation models. Each model represents a different function approximation architecture with varying number and form of basis functions. For example, one may assign a new model for each potential bandwidth in RBF-based basis functions.

The algorithm collects samples according to some behavior/exploration policy. Afterwards the samples are divided into two disjoint subsets $\mathcal{D}_{\text{learn}}$ and $\mathcal{D}_{\text{eval}}$. The former is used for the LSTD procedure while the latter is used for BERMIN. Our suggestion is that the size of these two sample sets should be in the same order.

In the main loop of Algorithm 4, it approximately evaluates a given policy π for all P sets of basis functions. This is different from the conventional LSPI in which we only deal with one set of basis functions. The next step is calling BERMIN and passing $\{Q_k\}_{k=1,2,\ldots,P}$ alongside the evaluation data set D_{eval}, the regression algorithm REGRESS$(\cdot)$, and the parameters δ/K, a, B, and τ. The regression algorithm can be any standard regression algorithm of choice, see Section 7.3.1 for some examples. The parameters a and B are defined as in Section 7.3.2. The confidence parameter passed is δ/K, so that the probability of failure caused by BERMIN procedure over all K iterations would be less than δ. Note that the value of δ is the probability of error caused only by BERMIN and not the error caused by LSTDQ.

Finally, there is a possibility of collecting new data samples after each iteration of LSPI. This new data may be collected according to the newly obtained policy π or any other

Algorithm 4 LSPI+Model Selection($\{\Phi_k\}_{k=1}^P, \gamma, \pi_0, \textsc{Regress}(\cdot), a, b, \delta, K$)

// $\{\Phi_k\}_{k=1}^P$: Sets of basis functions
// γ: Discount factor
// π_0: Initial policy
// Regress: The regression procedure
// a, B, τ: Parameters of BErMin
// K: The number of iterations
Collect data $\mathcal{D}$ in the form of $\{(X_i, A_i, R_i, X_i')\}$ following π_0 (or any other exploration policy)
Split $\mathcal{D}$ into $\mathcal{D}_{\text{learn}}$ and $\mathcal{D}_{\text{eval}}$
$\pi \leftarrow \pi_0$
for $t = 1$ to K **do**
  // Policy Evaluation step for all potential models
  for $k = 1, \ldots, P$ **do**
    $Q_k \leftarrow \text{LSTDQ}(\mathcal{D}_{\text{learn}}, \Phi_k, \gamma, \pi)$
  end for
  // Model Selection step
  $\hat{k} \leftarrow \textsc{BErMin}(\{Q_k\}_{k=1,2,\ldots,P}, \mathcal{D}_{\text{eval}}, \textsc{Regress}(\cdot), \frac{\delta}{K}, a, B, \tau)$
  // Policy Improvement step
  $\pi \leftarrow \hat{\pi}(\cdot; Q_{\hat{k}})$
  // [Optional step] Collecting more data
  [Optional] Collect data $\mathcal{D}^{(t)}$ according to π (or any other exploration policy).
  [Optional] Split $\mathcal{D}^{(t)}$ to $\mathcal{D}^{(t)}_{\text{learn}}$ and $D^{(t)}_{\text{eval}}$.
  [Optional] $\mathcal{D}_{\text{learn}} \leftarrow \mathcal{D}_{\text{learn}} \cup \mathcal{D}^{(t)}_{\text{learn}}$, $\mathcal{D}_{\text{eval}} \leftarrow \mathcal{D}_{\text{eval}} \cup \mathcal{D}^{(t)}_{\text{eval}}$.
end for

exploration policy. The resulting data samples should again be split into two disjoint subsets of $\mathcal{D}_{\text{learn}}$ and $\mathcal{D}_{\text{eval}}$ and combined with the previous data sets. Of course, one may gradually eliminate "old" data samples to avoid severe distribution mismatch caused by the non-stationarity of the whole process.

7.4 Theoretical Analysis

The goal of this section is to provide a theoretical justification for the BErMin procedure. We start with a rather abstract complexity regularization-based model selection algorithm and its analysis in Section 7.4.1. The main result proven there (Theorem 7.1), which goes beyond the setting of reinforcement learning, will be the basis of our main result, Theorem 7.2, which is presented in Section 7.4.2. Theorem 7.2 shows that BErMin has an oracle-like behavior, in the sense that with high probability it selects the model with the minimum Bellman error up to a multiplicative constant and some additional terms that converge to zero. Finally, in Section 7.4.3, we introduce the concept of adaptivity and prove that the oracle-like behavior of BErMin leads to its adaptivity (Theorem 7.3).

7.4.1 A Generic Model-Selection Theorem

The theorem presented in this section concerns a generic complexity regularization-based model selection procedure. The theorem and its proof technique are similar to Theorem 3 of Bartlett et al. [2002]. The main difference to this previous work is that our result is stated for an abstract setting where we are concerned with selecting the minimum amongst a set of values measured in noise, whereas Bartlett et al. [2002] developed their result in a specific supervised learning setting. Further, we make the role of non-central tail inequalities needed for the risk estimators explicit. Finally, we prove another related result, which will be useful

for our later developments. Nevertheless, the main proof technique is essentially the same as used in the proof of Theorem 3 of Bartlett et al. [2002]. For further similar results on complexity regularization, see Barron [1991]; Lugosi and Wegkamp [2004].

Theorem 7.1 (Key Technical Model Selection Theorem)**.** *Consider two sequences of random variables, $L_k, \mathcal{R}_k$, $k = 1, 2, \dots$. Assume that there exist positive constants c_1, c_2, c_3, c_4 and $0 < a < 1$, such that for any $0 < \delta \leq 1$ and $k = 1, 2, \dots$, the random variables L_k, $\mathcal{R}_k$ satisfy*

$$\mathbb{P}\left\{(1-a)\mathcal{R}_k \geq L_k - \frac{1}{c_2}\ln\frac{c_1}{\delta}\right\} \geq 1-\delta\,, \tag{7.2}$$

$$\mathbb{P}\left\{\frac{1}{1+a}\mathcal{R}_k \leq \mathbb{E}\left[\mathcal{R}_k\right] + \frac{1}{c_4}\ln\frac{c_3}{\delta}\right\} \geq 1-\delta\,. \tag{7.3}$$

Let C_k $(k = 1, 2, \dots)$ be a deterministic sequence that satisfies

$$c_5 \triangleq \sum_{k\geq 1}\exp\left(-c_2(1-a)C_k\right) < \infty\,, \tag{7.4}$$

$$c_6 \triangleq \sum_{k\geq 1}\exp\left(-c_4\frac{1+2a}{1+a}C_k\right) < \infty, \tag{7.5}$$

and define $\hat{k}$ by

$$\hat{k} \leftarrow \underset{k\geq 1}{\mathit{argmin}}[\mathcal{R}_k + C_k]\,.$$

Then, the following hold true:
(A) For any $0 < \delta < 1$, with probability at least $1-\delta$, it holds that

$$L_{\hat{k}} < (1-a^2)\min_{k\geq 1}\{\mathbb{E}\left[\mathcal{R}_k\right] + 2C_k\} + \frac{\ln(\frac{2c_1c_5}{\delta})}{c_2} + \frac{(1-a^2)\ln(\frac{2c_3c_6}{\delta})}{c_4}\,.$$

(B) For any $\alpha > 0$,

$$L_{\hat{k}} \leq (1-a^2)\min_{k\geq 1}\{\mathbb{E}\left[\mathcal{R}_k\right] + 2C_k\} + \alpha$$

holds with probability at least $1 - \left\{c_1c_5\ \exp\left(-\frac{c_2\alpha}{2}\right) + c_3c_6\ \exp\left(-\frac{c_4\alpha}{2(1-a^2)}\right)\right\}$.

In a typical application of this theorem, L_k would be the loss associated to some candidate k (from a set of at most countable candidates) and the random variable $\mathcal{R}_k$ would be a tightly concentrated, inflated estimate of L_k so that $(1-a)\mathcal{R}_k$ is still an overestimate of L_k, as required by condition (7.2). The theorem then yields that the loss associated with the selected candidate is not much larger than constant times the minimum of the losses biased by the "small' quantities C_k. In the appendix we show that conditions (7.2)-(7.3) are always satisfied for a slightly inflated estimate of L_k that tightly concentrates around its mean.

Proof. Fix $0 < \delta_1, \delta_2 \leq 1$. We start by bounding the deviation $\Delta = L_{\hat{k}} - (1-a^2)\min_k\{\mathbb{E}\left[\mathcal{R}_k\right] + 2C_k\}$. By adding and subtracting $(1-a)\min_k(\mathcal{R}_k + C_k)$, we can decompose Δ into two terms as follows:

$$\Delta = \underbrace{\left(L_{\hat{k}} - (1-a)\min_k(\mathcal{R}_k + C_k)\right)}_{\Delta_1} + (1-a)\underbrace{\left(\min_k(\mathcal{R}_k + C_k) - (1+a)\min_k(\mathbb{E}\left[\mathcal{R}_k\right] + 2C_k)\right)}_{\Delta_2}\,.$$

To bound the first term of this sum, we use that $\min_k(\mathcal{R}_k + C_k) = \mathcal{R}_{\hat{k}} + C_{\hat{k}}$, which holds thanks to the definition $\hat{k}$. Thus, we have

$$\Delta_1 = L_{\hat{k}} - (1-a)(\mathcal{R}_{\hat{k}} + C_{\hat{k}}) \leq \max_k\left\{L_k - (1-a)(\mathcal{R}_k + C_k)\right\}.$$

Choose any $0 < \delta'_k \le 1$ such that $\sum_k \delta'_k = \delta_1$. By condition (7.2), with probability $1 - \delta_1$, the quantity on the right-hand side of the last inequality is upper bounded by

$$\max_k \left\{ \frac{1}{c_2} \ln \frac{c_1}{\delta'_k} - (1-a)C_k \right\}.$$

In particular, if we choose $\delta'_k = \delta_1 / c_5 \exp(-c_2(1-a)C_k)$, the argument of the maximum becomes $\frac{1}{c_2} \ln \frac{c_1}{\delta'_k} - (1-a)C_k = \frac{1}{c_2} \ln \frac{c_1 c_5}{\delta_1}$ and thus we get that

$$\Delta_1 \le \frac{1}{c_2} \ln \frac{c_1 c_5}{\delta_1}$$

holds with probability $1 - \delta_1$.

Now, using $\min_\theta f(\theta) - \min_\theta g(\theta) \le \max_\theta (f(\theta) - g(\theta))$, Δ_2 can be bounded by

$$\Delta_2 \le (1+a) \max_k \left(\frac{\mathcal{R}_k}{1+a} - \mathbb{E}\left[\mathcal{R}_k\right] - \frac{1+2a}{1+a} C_k \right).$$

By condition (7.3), for any $0 < \delta''_k \le 1$ such that $\sum_k \delta''_k = \delta_2$, it holds with probability $1 - \delta_2$ that the quantity on the right-hand side of the above inequality is upper bounded by

$$(1+a) \max_k \left(\frac{1}{c_4} \ln \frac{c_3}{\delta''_k} - \frac{1+2a}{1+a} C_k \right).$$

Choosing $\delta''_k = \delta_2 / c_6 \exp(-c_4 \frac{1+2a}{1+a} C_k)$, we get that $\frac{1}{c_4} \ln \frac{c_3}{\delta''_k} - \frac{1+2a}{1+a} C_k = \frac{1}{c_4} \ln \frac{c_3 c_6}{\delta_2}$, therefore, with probability $1 - \delta_2$,

$$\Delta_2 \le \frac{1+a}{c_4} \ln \frac{c_3 c_6}{\delta_2}.$$

Combining the inequalities obtained for Δ_1 and Δ_2, we get that with probability $1 - (\delta_1 + \delta_2)$,

$$\Delta \le \frac{1}{c_2} \ln \frac{c_1 c_5}{\delta_1} + \frac{1-a^2}{c_4} \ln \frac{c_3 c_6}{\delta_2}. \tag{7.6}$$

To show Part (A), fix $0 < \delta \le 1$. Using the definition of Δ and (7.6), by choosing $\delta_1 = \delta_2 = \delta/2$ we get Part (A). To prove Part (B), fix some $\alpha > 0$. Choosing $\delta_1 = c_1 c_5 \exp(-c_2 \alpha / 2)$, $\delta_2 = c_3 c_6 \exp(-c_4 \alpha / (2(1-a^2)))$, from (7.6) we get that with probability $1 - (\delta_1 + \delta_2)$ the inequality $\Delta \le \alpha$ holds, thus finishing the proof. □

7.4.2 Model Selection for Reinforcement Learning and Planning

In this section we state and prove our main result which shows that BERMIN has an oracle-like behavior. We prove the result under the following assumption.

Assumption A20 Assume that the following hold:

1. The standard offline sampling assumption is satisfied by the data set

   $$\mathcal{D}''_n = \{(X_1, A_1, R_1, X'_1), \ldots, (X_n, A_n, R_n, X'_n)\}$$

   and the time-homogeneous Markov chain $X_1, X_2, \ldots, X_n$ uniformly quickly forgets its past with a forgetting time τ (cf. Definition 7.1 in Appendix 7.B).

2. The functions Q_k, $\tilde{Q}_k$, $T^* Q_k$ ($k \ge 1$) are bounded by a deterministic quantity $B > 0$.

3. The functions Q_k ($k \ge 1$) are deterministic.

4. For each k and for any $0 < \delta'_k < 1$, $(\tilde{Q}_k, \bar{b}_k) = \textsc{Regress}(\mathcal{D}'_{m,k}, \delta'_k)$ are $\sigma(\mathcal{D}'_m)$-measurable, $\bar{b}_k \in [0, 4B^2]$ and $\|\tilde{Q}_k - T^* Q_k\|_\nu^2 \le \bar{b}_k$ holds with probability at least $1 - \delta'_k$.

5. For $(X_i, A_i, R_i, X'_i) \in \mathcal{D}''_n$, the distribution of (X_i, A_i) given $\mathcal{D}'_m$ is ν: $\mathbb{P}\{(X_i, A_i) \in U | \mathcal{D}'_m\} = \nu(U)$ for any measurable set $U \subset \mathcal{X} \times \mathcal{A}$.

A couple of remarks on these assumptions are in order.

Remark 7.4. The standard offline sampling assumption was discussed in Section 2.2.1. The additional assumption here demands that the Markov chain should "forget its past" uniformly fast. The actual definition, which we think is often satisfied, is somewhat technical and is given in the appendix. Here we note that this condition is satisfied if the Markov chain is uniformly ergodic (or, in other words, if the so-called Doeblin condition holds for the Markov chain [Meyn and Tweedie, 2009]). Note that if the chain mixes but the "mixing rate" is slow, a result similar to the one presented below would still hold, but possibly with a worse rate. On another note, although we have not made any specific distributional assumptions about $\mathcal{D}'_m$, it is expected that $\mathcal{D}'_m$ should satisfy similar assumptions to $\mathcal{D}''_n$ to make $\bar{b}_k$ small.

Remark 7.5. If the immediate rewards are bounded with probability one, most algorithms would return deterministically bounded value functions. If this is not known to hold for some algorithm, but a bound $r_{\max}$ on the immediate reward function is known, then boundedness can be achieved by truncating the value functions Q_k and $\tilde{Q}_k$ so that they take values in the interval $[-B, B] = [-r_{\max}/(1-\gamma), r_{\max}/(1-\gamma)]$ (i.e., instead of $Q_k(x,a)$, use $\min(\max(Q_k(x,a), -B), B)$). Since the target of learning in both cases is a function with range contained in $[-B, B]$, truncating the action-values this way introduces no loss of quality.

Remark 7.6. That the functions (Q_k) are deterministic is not an essential requirement, as already noted in Remark 7.1.

Remark 7.7. In Line 5 of Algorithm 3, we call $(\tilde{Q}_k, \bar{b}_k) \leftarrow \textsc{Regress}(\mathcal{D}'_{m,k}, \delta'_k)$. The condition that $\sum_{k \geq 1} \delta'_k = \delta/2$ ensures that simultaneously, for all $k \geq 1$, $\|\tilde{Q}_k - T^* Q_k\|_\nu^2 \leq \bar{b}_k(\delta'_k)$ holds with probability at least $1 - \delta/2$.

Remark 7.8. One approach to get the required high probability estimates $\bar{b}_k$ is described in Section 7.D.

Remark 7.9. The success of BErMin will depend critically on the quality of the regression procedure, Regress, that it calls. If the value-function estimation procedure A used to calculate the candidate action-value functions is available, one appealing idea is to reuse this procedure for the purpose of computing the functions $(\tilde{Q}_k)$. This can be done when A also accepts the value of the discount factor as input γ. In this case, one could feed A with $\gamma = 0$ and the data

$$\mathcal{D}'_{m,k} = \left\{ \left(X, A, (\hat{T}^* Q_k)(X, A), X' \right) \,:\, (X, A, R, X') \in \mathcal{D}'_m \right\}$$

to produce $\tilde{Q}_k$, where we have replaced the immediate rewards in the data with the estimates of $T^* Q_k$.[2] This works because with $\gamma = 0$ the problem of finding the optimal value function becomes equivalent to estimating the immediate reward function based on the available sample. When producing the estimate $\tilde{Q}_k$ it would make sense to use the same tuning of A as the one used to produce Q_k. This will be further explored in Section 7.4.3. Nevertheless, one is not limited to this choice and, in fact, it makes perfect sense to use an adaptive regression procedure. This can be done based on Theorem 7.1 or in many other ways (for some recent works on adaptive regression estimation, refer to e.g., Wegkamp 2003; van der Vaart et al. 2006 or Arlot and Celisse 2009).

We are ready to present the main result of this work:

[2] Note that here and in what follows we use the notation $\hat{T}^*$ liberally to be interpreted based on the local context as the empirical Bellman operator underlying the dataset whose samples $\hat{T}^*$ interacts with in the given expression. Thus, in the above case, $(\hat{T}^* Q)(X, A)$ is meant to be computed based on $\mathcal{D}'_m$.

Theorem 7.2 (Model Selection for RL/Planning). *Let Assumption A20 hold. Consider the* BERMIN *algorithm defined in Section 7.3 used with some* $0 < a < 1$, $0 < \delta \le 1$, *and* $(C_k)_{k\ge 1}$ *such that*

$$S \triangleq \sum_{k\ge 1} \exp\left(-\frac{(1-a)^2 a\, n}{16B^2\tau\,(1+a)} C_k\right) < \infty \tag{7.7}$$

holds. Let $\hat{k}$ *be the index selected by* BERMIN. *Then, with probability at least* $1-\delta$,

$$\|Q_{\hat{k}} - T^* Q_{\hat{k}}\|_\nu^2 \le 4(1+a) \min_{k\ge 1}\left\{ \frac{2}{(1-a)^2}\|Q_k - T^* Q_k\|_\nu^2 + \frac{3}{(1-a)^2}\bar{b}_k + 2C_k \right\} + \frac{96B^2\tau\,(1+a)}{(1-a)^2 a\, n} \ln\left(\frac{4S}{\delta}\right).$$

Note that $C_k = \frac{32B^2\tau(1+a)}{(1-a)^2 an}\ln(k)$ satisfies $S < \infty$ (in particular, with this choice we get $S = \pi^2/6$). A detailed discussion of the result is given after its proof.

Proof. By the triangle inequality and $(|x|+|y|)^2 \le 2(x^2+y^2)$, we get

$$\|Q_{\hat{k}} - T^* Q_{\hat{k}}\|_\nu^2 \le 2\left(\|Q_{\hat{k}} - \tilde{Q}_{\hat{k}}\|_\nu^2 + \|\tilde{Q}_{\hat{k}} - T^* Q_{\hat{k}}\|_\nu^2\right).$$

Define $L_k = \|\tilde{Q}_k - Q_k\|_\nu^2 + (1-a)\bar{b}_k$. The first term on the right-hand side of the last inequality can be upper bounded by $L_{\hat{k}}$, while, outside of an error event $\mathcal{E}_1$ of probability mass at most $\delta/2$, the second term can be upper bounded by $\bar{b}_{\hat{k}}$. Using the definition of L_k, we can further upper bound this term by $L_{\hat{k}}/(1-a)$, thus obtaining that on $\mathcal{E}_1^c$

$$\|Q_{\hat{k}} - T^* Q_{\hat{k}}\|_\nu^2 \le \frac{2(2-a)}{1-a} L_{\hat{k}} \le \frac{4}{1-a} L_{\hat{k}}\,.$$

Thus, the problem is reduced to that of bounding $L_{\hat{k}}$. For this, we will use Theorem 7.1. Let

$$\|\tilde{Q}_k - Q_k\|_n^2 = \frac{1}{n}\sum_{(x,a,r,x')\in\mathcal{D}_n''} (\tilde{Q}_k(x,a) - Q_k(x,a))^2.$$

Note that by our assumptions and conventions for multisets, this sum has n terms. Define

$$\mathcal{R}_k = \frac{1}{(1-a)^2}\|\tilde{Q}_k - Q_k\|_n^2 + \bar{b}_k\,.$$

With these definitions, the index $\hat{k}$ returned by BERMIN can be given as

$$\hat{k} = \operatorname*{argmin}_{k\ge 1}\left[\mathcal{R}_k + C_k\right].$$

Thus, provided that $(\mathcal{R}_k)$, (L_k) satisfy (7.2)–(7.3) and (C_k) satisfies (7.4)–(7.5), we will be able to conclude from Theorem 7.1 a bound on $L_{\hat{k}}$ and thus also on the Bellman error of the selected action-value function. Since $\tilde{Q}_k$, $\bar{b}_k$ are themselves a function of $\mathcal{D}_m'$, we will use Theorem 7.1 on the probability space $\Omega_m = (\Omega, \sigma_\Omega, \mathbb{P}_m)$ with $\mathbb{P}_m(\cdot) = \mathbb{P}(\cdot|\mathcal{D}_m')$, i.e., we will apply the theorem on the probability space obtained by conditioning on $\mathcal{D}_m'$. Since a bound on a conditional probability gives a bound on the unconditioned probability, this will be sufficient to conclude a high probability bound on $L_{\hat{k}}$.

Let us consider (7.2). This condition requires that for some $c_1, c_2 > 0$, for any $0 < \delta' \le 1$, $\mathbb{P}_m(L_k - (1-a)\mathcal{R}_k \le \frac{1}{c_2}\ln\frac{c_1}{\delta'}) \ge 1-\delta'$. By the definition of L_k and $\mathcal{R}_k$,

$$\begin{aligned} L_k - (1-a)\mathcal{R}_k &= \|\tilde{Q}_k - Q_k\|_\nu^2 + (1-a)\bar{b}_k - \left(\frac{1}{1-a}\|\tilde{Q}_k - Q_k\|_n^2 + (1-a)\bar{b}_k\right) \\ &= \|\tilde{Q}_k - Q_k\|_\nu^2 - \frac{1}{1-a}\|\tilde{Q}_k - Q_k\|_n^2\,. \end{aligned}$$

Our plan is to use Lemma 7.7 of Appendix 7.C to provide the required bound. For this notice that $\mathbb{E}\left[\|\tilde{Q}_k - Q_k\|_n^2 | \mathcal{D}'_m\right] = \|\tilde{Q}_k - Q_k\|_\nu^2$ and that $\|\tilde{Q}_k - Q_k\|_n^2$ can be written as an average of the values taken by the function $f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, $(x, a) \mapsto (\tilde{Q}_k(x, a) - Q_k(x, a))^2$ over a Markov chain taking values in $\mathcal{X} \times \mathcal{A}$. By Assumption A20.1, the forgetting time of the underlying $\mathcal{X}$-valued chain is bounded by τ. It follows from the definition of forgetting times and that the actions are sampled from a fixed behavior policy that the forgetting time of the $\mathcal{X} \times \mathcal{A}$-valued chain is also bounded by τ. Further, by Assumption A20.2, the range of f is in $[0, 4B^2]$. Thus, by the first part of Lemma 7.7, $\mathbb{P}_m(\|\tilde{Q}_k - Q_k\|_\nu^2 - \frac{1}{1-a}\|\tilde{Q}_k - Q_k\|_n^2 \le \frac{8B^2(1+a)\tau}{(1-a)an} \ln \frac{1}{\delta'}) \ge 1 - \delta'$. Thus, condition (7.2) holds with $c_1 = 1$ and $c_2 = \frac{(1-a)an}{8B^2(1+a)\tau}$.

Now, let us consider (7.3). This condition requires that for some $c_3, c_4 > 0$, for each $0 < \delta' \le 1$, $\mathbb{P}_m(\frac{1}{1+a}\mathcal{R}_k - \mathbb{E}[\mathcal{R}_k | \mathcal{D}'_m] \le \frac{1}{c_4} \ln \frac{c_3}{\delta'}) \ge 1 - \delta'$. Again, $\mathcal{R}_k$ is an average of the function $f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, $(x, a) \mapsto \frac{1}{(1-a)^2}(\tilde{Q}_k(x, a) - Q_k(x, a))^2 + \bar{b}_k$ over an $\mathcal{X} \times \mathcal{A}$-valued Markov chain with forgetting time bounded by τ. The range of function f is contained in $[0, 4B^2(1 + \frac{1}{(1-a)^2})]$. Therefore, the second part of Lemma 7.7 gives that the required inequality holds with $c_3 = 1$, $c_4 = \frac{(1-a)^2 an}{8B^2(1+(1-a)^2)\tau}$.

It remains to check (7.4) and (7.5). A simple calculation gives that condition (7.7) ensures that both $c_5 = \sum_{k \ge 1} \exp(-c_2(1-a)C_k)$ and $c_6 = \sum_{k \ge 1} \exp(-c_4 \frac{1+2a}{1+a} C_k)$ are finite and upper bounded by S. Therefore, by Part (A) of Theorem 7.1,

$$L_{\hat{k}} \le (1 - a^2) \min_{k \ge 1} \left[\frac{1}{(1-a)^2} \|\tilde{Q}_k - Q_k\|_\nu^2 + \bar{b}_k + 2C_k \right] + \Delta_1 \,, \tag{7.8}$$

holds outside of an error event $\mathcal{E}_2$ of probability mass at most $\delta/2$, where

$$\Delta_1 = \frac{\ln(\frac{2c_5}{\delta/2})}{c_2} + \frac{(1-a^2)\ln(\frac{2c_6}{\delta/2})}{c_4} \le \frac{8B^2\tau(1+a)(2+(1-a)^2)}{(1-a)a\,n} \ln\left(\frac{4S}{\delta}\right) .$$

It remains to upper bound $\|\tilde{Q}_k - Q_k\|_\nu^2$. For this note that on $\mathcal{E}_1^c$ the inequalities $\|\tilde{Q}_k - T^*Q_k\|_\nu^2 \le \bar{b}_k$ hold simultaneously for all $k \ge 1$. Hence, on this event, $\|\tilde{Q}_k - Q_k\|_\nu^2 \le 2(\|Q_k - T^*Q_k\|_\nu^2 + \bar{b}_k)$. Thus, on $(\mathcal{E}_1 \cup \mathcal{E}_2)^c$,

$$\|Q_{\hat{k}} - T^*Q_{\hat{k}}\|_\nu^2 \le 4(1+a) \min_{k \ge 1} \left[\frac{2}{(1-a)^2} \|Q_k - T^*Q_k\|_\nu^2 + \frac{2+(1-a)^2}{(1-a)^2} \bar{b}_k + 2C_k \right] + \frac{4\Delta_1}{1-a}$$

Bounding $2 + (1-a)^2$ by 3 gives the final result. □

To gain a better understanding of the bound of Theorem 7.2, we discuss the contribution of each of its right-hand side terms.

The term $\|Q_k - T^*Q_k\|_\nu^2$ is the true Bellman error of each candidate action-value function Q_k, and is a measure of the approximation error. This is the main quantity of interest and the ultimate goal of the minimization, which is not accessible to us. An oracle, having access to T^*Q_k, would select $\hat{k} = \operatorname{argmin}_{k \ge 1} \|Q_k - T^*Q_k\|_\nu^2$.

By definition, the term $\bar{b}_k$ is a bound on how well $\tilde{Q}_k$ approximates T^*Q_k. We need two conditions to hold true to make this term small: The regression procedure Regress should return a good estimate of T^*Q_k, while the bound returned on the excess risk by the same procedure should also be a tight bound on the excess-risk of the returned regressor. In Section 7.D of the Appendix we show how these goals can be achieved by building on Theorem 7.1 in a quite general situation. To make the whole procedure competitive with an oracle, one should ensure that $\bar{b}_k$ is comparable to the size of Bellman-error $\|Q_k - T^*Q_k\|_\nu^2$. How to achieve this is further discussed in Section 7.4.3.

The third term of the bound is the complexity regularizer C_k and shows the price we pay to have an algorithm that works with a very large (or even infinite) number of models. As discussed earlier, the choice of C_k should reflect our prior belief about the suitability of

the candidates. Note that if one has a finite number of models, then one can use $C_k = 0$. In the general case, C_k will depend on k, but it is still expected to be small compared to the other terms. The complexity regularizer has an information theoretic interpretation, which is discussed by Barron [1991]; Barron et al. [2008].

The term outside the minimizer comes from the randomness of the sample $\mathcal{D}_n''$ used to estimate one component of the Bellman error. This term, just like C_k, converges to zero at a parametric rate, and it is thus expected to be small compared to the other terms. Note the tradeoff between the terms (C_k) and this last term.

Another tradeoff exists between the first two and the last two terms. This tradeoff is governed by a: as a approaches zero, the constant in front of the first two terms become smaller, but the last two terms diverge to infinity (see the specific form of C_k after the statement of the theorem). Moreover, as a approaches 1, the multipliers of all these terms blow up. As the first two terms often go to zero slower than the last one as the number of samples grows, one expects that a value of a close to zero will give the best tradeoff and in fact letting a go to zero like $a \sim n^{-\frac{1}{2}}$ might be the best choice. However, when the first two terms are fast (i.e., they converge to zero at the $O(1/n)$ rate) then one should keep a bounded away from zero to get the best asymptotic rate.

Remark 7.10. The result also holds true for policy evaluation, when given some policy π, the goal is to select a function Q_k that minimizes the Bellman error $\|T^\pi Q_k - Q_k\|_\nu$. In order to use BErMin for this problem, in the definition of the dataset, $\hat{T}^\pi$ should be used in place of $\hat{T}^*$. In fact, the only property of $\hat{T}^*$ that we used in the proof was the property stated in Proposition 2.1, which holds for both T^* and T^π.

Remark 7.11. If the forgetting time τ or an upper bound thereof is not known, one may use $\hat{\tau}(n) = \tau_0\, f(n)$ in the BErMin procedure for some $\tau_0 > 0$, and a positive-valued function f that diverges. Then, as soon as $\hat{\tau} > \tau$, the conclusion of Theorem 7.2 will hold with τ in the bound replaced by $\hat{\tau}$. In order to get the asymptotically best rate, one should choose a function f that grows slowly and a small value of τ_0. For example, when $f(n) = \ln(n)$, the asymptotic bound is increased only by a logarithmic factor. However, a slowly growing f with a small τ_0 can lead to a poor transient performance. On the other hand, if f grows faster (e.g., $f(n) = n^r$ for some $0 < r < 1$) or when τ_0 is larger, the transient performance is expected to improve at the price of a worse asymptotic performance.

7.4.3 Adaptivity

The purpose of this section is to show that BErMin can be made an adaptive procedure in a well-defined sense. We start with explaining what we mean by adaptivity.

The concept of adaptivity

We consider the special case when the algorithm A used to compute Q_k, in addition to a dataset, takes as input a function space $\mathcal{F}(p_k)$, the discount factor $0 < \gamma < 1$ and the confidence parameter $0 < \delta \le 1$. The idea is that when A is run with this input, it will output an action-value function belonging to $\mathcal{F}(p_k)$. For a given k, $\mathcal{F}(p_k)$ may or may not hold the optimal action-value function. As a result, $\mathcal{F}(p_k)$ will impact the quality of Q_k returned by A in two ways: First, if $\mathcal{F}(p_k)$ is large, the limiting Bellman error of Q_k (as the number of samples converges to infinity) is expected to be smaller. Let us denote this quantity by $a_k(T^*)$ (the parameters signifying that the limiting error depends on k and on the MDP through T^*). The second effect is that if $\mathcal{F}(p_k)$ is large, the algorithm A will be more susceptible to overfitting. Overall, we expect that for any k, T^*, $0 < \delta \le 1$, $n \ge 1$, a high-probability bound of the form

$$\|Q_k - T^* Q_k\|_\nu^2 \le a_k(T^*) + c_{T^*}\, b_k(n, \ln(1/\delta)), \tag{7.9}$$

which holds with probability at least $1-\frac{2}{\pi^2}\delta$, will hold for A.[3] Here, the second term bounds the error that results from using a finite number of samples. In this term, c_{T^*} is a constant that depends on T^* only (i.e., on the MDP), but is independent of $\mathcal{F}(p_k)$, n and δ. On the other hand, b_k does not depend on T^*. If in the limit of an infinite sample, $\|Q_k - T^*Q_k\|_\nu^2$ converges to $\inf_{Q\in\mathcal{F}(p_k)} \|Q - T^*Q\|_\nu$, then $a_k(T^*) = \inf_{Q\in\mathcal{F}(p_k)} \|Q - T^*Q\|_\nu^2$. Thus, in this case $a_k(T^*)$ becomes equal to the (squared) *approximation error* underlying $\mathcal{F}(p_k)$ and the second term is said to bound the *estimation error*. Typically, b_k is a polynomial of the ratio of its arguments and scales with how "large" $\mathcal{F}(p_k)$ is and it is expected that $b_k \to \infty$ as $k \to \infty$. It is assumed that (7.9) is a tight bound of this form (at this stage, the particular form of the above bound is unimportant). Note that being a tight bound, in general one cannot compute this bound as this would require *a priori* knowledge of quantities which, in general, are *a priori* unknown. For example, $a_k(T^*)$ is typically unknown. Thus, only an oracle could evaluate these bounds.

By (7.9), it follows that the inequalities

$$\|Q_k - T^*Q_k\|_\nu^2 \le a_k(T^*) + c_{T^*}\, b_k(n, \ln(k^2/\delta)) \tag{7.10}$$

hold *simultaneously* for all $k \ge 1$, with probability at least $1-\delta/3$. Thus, an oracle, having access to the bounds on the right-hand side could select the index k^* such that $\|Q_{k^*} - T^*Q_{k^*}\|_\nu^2 = \beta_n$, where

$$\beta_n \triangleq \min_{k\ge 1} \left\{ a_k(T^*) + c_{T^*}\, b_k(n, \ln(k^2/\delta)) \right\}.$$

We call a procedure *adaptive* if it only uses data set $\mathcal{D}_n$ but still matches the error of the candidate k^* up to a constant factor. Formally, if $\hat{k}$ is the index selected by a procedure then we call the procedure adaptive, if for some $C, c \ge 1$ it holds that for each MDP of interest[4], $n \ge 1$, $0 < \delta < 1/c$, we have

$$\left\|Q_{\hat{k}} - T^*Q_{\hat{k}}\right\|_\nu^2 \le C \min_{k\ge 1} \left\{ a_k(T^*) + c_{T^*} b_k(n, \ln(k^2/\delta)) \right\},$$

with probability $1 - c\,\delta$.

The adaptivity of BErMin

In this section we assume that $m = n$, i.e., the initial data has an even length which is split into two equal halves. The purpose of this section is to show that BErMin can be used as the basis of an adaptive procedure. For this, we propose to use A as the regression procedure Regress used in BErMin. To make our proposal formal, assume that A takes four parameters: the function space, the dataset, the discount factor, and the confidence parameter, and it returns both an action-value estimate and a confidence bound. We propose that BErMin should use

$$(\tilde{Q}_k, \bar{b}_{k,n}(\delta)) = \mathsf{A}\left(\mathcal{F}(p_k), \mathcal{D}'_{n,k}, 0, \tfrac{2}{\pi^2}\tfrac{\delta}{k^2}\right)$$

with

$$\mathcal{D}'_{n,k} = \mathcal{D}'_n(Q_k) = \left\{ \left(X_1, A_1, (\hat{T}^*Q_k)(X_1, A_1), X'_1\right), \ldots, \left(X_n, A_n, (\hat{T}^*Q_k)(X_n, A_n), X'_n\right) \right\}.$$

Since $\gamma = 0$, algorithm A acts as a regression procedure that works in the function space $\mathcal{F}(p_k)$ (and will in fact disregard the next states $X'_1, \ldots, X'_n$).

We make the following assumption on $\bar{b}_{k,n}$ returned by A:

[3] The purpose of constant $\frac{2}{\pi^2}$ is to simplify subsequent developments, but is otherwise unimportant due to the logarithmic dependence of b_k on $1/\delta$.

[4] The class of MDPs can be restricted. Then the procedure is called adaptive within the chosen class.

Assumption A21 Tightness of $\bar{b}_{k,n}$ There exists some $C \geq 1$ such that for each MDP of interest, sample-size n, model index k, action-value function Q bounded by B and confidence parameter $0 < \delta < 1$, when A is fed with $\mathcal{F}(p_k)$, $\mathcal{D}'_n(Q)$, $\gamma = 0$, and δ then $\bar{b}_{k,n}(\delta)$ returned by A satisfies

$$\bar{b}_{k,n}(\delta) \leq C \left[\inf_{Q' \in \mathcal{F}(p)} \|Q' - T^* Q\|_\nu^2 + b_k \left(n, \ln(\tfrac{2}{\pi^2 \delta})\right) \right] \tag{7.11}$$

with probability at least $1 - \delta$.[5]

Note that we make no assumption on how A behaves when its input is different from the above. In particular, we make no assumption about whether $\bar{b}_{k,n}(\delta)$ will be tight when A is fed with $\gamma > 0$. A crucial point about the above assumption is that it uses the same b_k functions which are used in the definition of adaptivity.

Since $\bar{b}_{k,n}(\delta)$ is an upper bound on the error of the action-value function returned by A, the above assumption implies two things about A when used as a regression procedure. First, in the limit of infinite samples the function returned should become close (up to a positive constant) to the theoretically best approximation error. In fact, many regression algorithms (such as the ones mentioned earlier) satisfy this condition (and can in fact achieve the approximation error). Second, the term bounding estimation error underlying A when used as a regression procedure, apart from a constant factor, should be the same as the corresponding term when A is used to approximate the fixed point of some non-constant operator. This is again reasonable, since regression in general is expected to be easier than fixed point estimation.

Now, we are ready to state the main result of this section:

Theorem 7.3. *Let Assumptions A20 and A21 hold and assume that $m = n$. In addition, assume that* (i) *for each $k \geq 1$, (7.9) holds with probability at least $1 - \delta$ where $c_{T^*} \geq C^*$ for some positive constant C^* that is independent of T^*; and* (ii) *for any index $k \geq 1$, $L > 0$, we have $b_k(n, L) = \Omega(L/n)$. Then, when* BERMIN *is used with* REGRESS $= \mathsf{A}$ *with $\gamma = 0$ and $C_k = \frac{32 B^2 \tau(1+a)}{(1-a)^2 a n} \ln(k)$, the resulting procedure is adaptive: there exists a positive constant C'' such that for each MDP, $n \geq 1$, and $0 < \delta < 1$, the Bellman-error of the action-value function selected by* BERMIN *is bounded by*

$$\left\|Q_{\hat{k}} - T^* Q_{\hat{k}}\right\|_\nu^2 \leq C'' \beta_n = C'' \min_{k \geq 1} \left[a_k(T^*) + c_{T^*} \, b_k \left(n, \ln(\tfrac{k^2}{\delta})\right) \right],$$

with probability at least $1 - \frac{5}{3}\delta$.

Proof. From Theorem 7.2, with the choice of $C_k = \frac{32 B^2 \tau(1+a)}{(1-a)^2 a n} \ln(k)$, we have that with probability at least $1 - \delta$,

$$\left\|Q_{\hat{k}} - T^* Q_{\hat{k}}\right\|_\nu^2 \leq \min_{k \geq 1} \left[c_1 \, \|Q_k - T^* Q_k\|_\nu^2 + c_2 \, \bar{b}_{k,n} \left(\tfrac{2}{\pi^2} \tfrac{\delta}{k^2}\right) + c_3 \frac{\ln(k)}{n} \right] + c_4 \frac{\ln(1/\delta)}{n}$$

holds for some constants $c_1, c_2, c_3, c_4 > 0$ which do not depend on the MDP, δ and n. From Assumption A21, we get that the inequalities

$$\begin{aligned} \bar{b}_{k,n} \left(\tfrac{2}{\pi^2} \tfrac{\delta}{k^2}\right) &\leq C \left[\inf_{Q' \in \mathcal{F}(p)} \|Q' - T^* Q_k\|_\nu^2 + b_k(n, \ln(k^2/\delta)) \right] \\ &\leq C \left[\|Q_k - T^* Q_k\|_\nu^2 + b_k(n, \ln(k^2/\delta)) \right] \end{aligned} \tag{7.12}$$

hold simultaneously for all $k \geq 1$ with probability at least $1 - \delta/3$. Thus, for some $c'_1, c'_2 > 0$,

$$\left\|Q_{\hat{k}} - T^* Q_{\hat{k}}\right\|_\nu^2 \leq \min_{k \geq 1} \left[c'_1 \, \|Q_k - T^* Q_k\|_\nu^2 + c'_2 \, b_k \left(n, \ln(\tfrac{k^2}{\delta})\right) + c_3 \frac{\ln(k)}{n} \right] + c_4 \frac{\ln(1/\delta)}{n}$$

[5] As before, the constant $\pi^2/2$ is included only to simplify some further results.

holds with probability at least $1-\frac{4}{3}\delta$. Now, by *(ii)*, $\frac{\ln(k)}{n} = \mathcal{O}\left(b_k\left(n, \ln\frac{k^2}{\delta}\right)\right)$ and $\frac{\ln(1/\delta)}{n} = \mathcal{O}\left(b_k\left(n, \ln\frac{k^2}{\delta}\right)\right)$. Hence, with some $C' > 0$, on the event where the previous inequality holds,

$$\left\|Q_{\hat{k}} - T^*Q_{\hat{k}}\right\|_\nu^2 \le C' \min_{k\ge 1}\left[\left\|Q_k - T^*Q_k\right\|_\nu^2 + b_k\left(n, \ln\frac{k^2}{\delta}\right)\right]$$

holds, too. By (7.9), the inequalities

$$\left\|Q_k - T^*Q_k\right\|_\nu^2 \le a_k(T^*) + c_{T^*}\, b_k(n, \ln(k^2/\delta))$$

hold simultaneously for all $k \ge 1$ with probability $1-\delta/3$. Hence, with probability $1-\frac{5}{3}\delta$, with some $C'' > 0$,

$$\left\|Q_{\hat{k}} - T^*Q_{\hat{k}}\right\|_\nu^2 \le C'' \min_{k\ge 1}\left[a_k(T^*) + c_{T^*}\, b_k\left(n, \ln(\tfrac{k^2}{\delta})\right)\right] = C''\beta_n\,,$$

where we used that, by assumption, c_{T^*} is bounded away from zero. □

7.5 Conclusion

In this work we suggested a principled approach for the tuning of reinforcement learning algorithms in the offline and non-interactive scenario. The problem was formulated as that of finding an action-value function with a small Bellman error among a set of candidate functions. BErMin, a complexity regularization-based algorithm, was introduced for this purpose.

Our main theoretical result, Theorem 7.2, is a finite-sample high-probability upper bound that shows that the Bellman error of the action-value function selected by BErMin is almost as small as that of an oracle who has access to the true Bellman errors. This result was further elaborated in Section 7.4.3, where we have shown that BErMin can be made adaptive in the sense that it can compete with an oracle who selects the model with the smallest error bounds (Theorem 7.3). As far as we know, this is the first work that considers adaptivity in a reinforcement learning scenario. The main message of our results is that just like in supervised learning, it is possible to learn almost as fast as if one had extra *a priori* information.

In this chapter we focused on the goal of finding an action-value function with a small Bellman error. However, the primary goal in reinforcement learning is to find good policies. Is it possible to derive results similar to ours for this alternative problem? In what follows we consider two possible approaches.

First, still sticking to the action-value based approach, one might be tempted to consider the *projected Bellman error*, instead of the Bellman error. To recap, for some function space $\mathcal{F}^{|\mathcal{A}|}$, the projected Bellman error of $Q \in \mathcal{F}^{|\mathcal{A}|}$ is defined as $\|Q - \Pi_{\mathcal{F}^{|\mathcal{A}|}} T^*Q\|$, where $\Pi_{\mathcal{F}^{|\mathcal{A}|}}$ is the projection operator that maps its argument to the closest point on $\mathcal{F}^{|\mathcal{A}|}$ w.r.t. an appropriate norm. The projected Bellman error is typically defined for linear function spaces $\mathcal{F}^{|\mathcal{A}|}$, therefore we also restrict our discussion to such spaces. The advantage of the projected Bellman error then is that its magnitude can be readily estimated based on a sample (see, e.g., Antos et al. 2008b; Szepesvári 2010). However promising this is, unfortunately, the projected Bellman error is unsuitable for model selection purposes as it eliminates the component of the error that is orthogonal to $\mathcal{F}^{|\mathcal{A}|}$ (see also Remark 3.1 in Section 3.2). Thus, even if one could calculate the exact values of the projected Bellman error, this knowledge would be useless for model selection purposes. This limitation of the projected Bellman error is also apparent if we note that under the so-called on-policy sampling condition and when $\mathcal{F}^{|\mathcal{A}|}$ is a nontrivial space, the projected Bellman error is always zero, independently of the choice of $\mathcal{F}^{|\mathcal{A}|}$. Therefore, the projected Bellman-error alone contains no information about the suitability of $\mathcal{F}^{|\mathcal{A}|}$.

Let us consider the next alternative, which we might call model-(or simulation-)based policy selection. Assume as before that the problem is already reduced to that of selecting the best policy from a list of policy candidates $\pi_1, \ldots, \pi_P$. Let the performance be measured as the expected total discounted reward with respect to some known initial distribution ρ. For an MDP M and policy π, let this measure be $V^\pi(M, \rho)$.

One way to avoid using value functions is to use part of the data to build an approximate model $\hat{M} = (\mathcal{X}, \mathcal{A}, \hat{P}, \gamma)$ of the MDP of interest. Assume that for any learned model $\hat{P}$, one can efficiently generate *virtual* trajectories for the initial distribution ρ and any policy of interest π. For $1 \leq i \leq P$, let $V^{\pi_i}(\hat{M}, \rho)$ be the average of the returns obtained by following policy π_i in $\hat{M}$. If $\hat{P}$ is close enough to P, in an appropriate norm, and enough virtual trajectories are used, the estimates of $V^{\pi_i}(\hat{M}, \rho)$ will be close to $V^{\pi_i}(M, \rho)$ and thus it makes sense to select the policy with the maximum estimated expected return. The quality of this procedure will ultimately depend on how well $\hat{M}$ approximates M (since generating virtual trajectories is cheap), i.e., the problem of designing an effective policy selection method is reduced to that of learning a good generative model. Model learning based on sampled transitions falls into the realm of supervised learning. Hence, having an adaptive procedure for policy-selection will hinge upon if we have an adaptive model-learning procedure. Studying the advantages and disadvantages of this approach will be the topic of future work.

Future Work

Although in this chapter we made some progress toward reinforcement learning algorithms that require minimum human supervision, the problem is far from being solved. In particular, the following questions require further investigation:

- How to generate the list of candidate action-value functions $(Q_1, Q_2, \ldots)$? In what order should we run the methods available? We briefly discussed this issue in Remark 7.3 in an abstract setting. However, a more thorough, systematic approach would be desired and much remains to be done in this respect.

- How can one achieve adaptivity in online and interactive learning scenarios? The current work is specific to the offline learning scenario, where we could not benefit from interacting with the environment. However, it is unclear if adaptivity can be achieved in the online, interactive scenarios.

- How can one construct data-dependent estimates of the forgetting time parameter τ? Both Meir [2000] and Modha and Masry [1998] face a similar situation; their respective procedures require the knowledge of the β-mixing coefficients of the dependent stochastic process. As far as we know, there is yet no rigorous procedure to estimate such parameters in the general case. Nevertheless, McDonald [2010] has recently proposed to use a mutual information-based estimator to upper bound the β-mixing coefficients, but the sample-efficiency of the method is yet to be shown. Meanwhile, one may use the procedure described in Remark 7.11 at the cost of a marginally slower than $1/n$ extra loss.

- What is the relation between the quality of the solution of the fixed point of the Bellman optimality operator and the performance of the corresponding greedy policy? An extension of Theorem 5.3 of Munos 2007 alongside the machinery developed in Chapter 3 could be helpful in this respect.

- We derived some data-dependent bounds on the excess-risk of a regression procedure that operates in a large function space which suited our immediate needs. However, the bound is asymptotic in nature and is potentially suboptimal. Can this bound be improved?

- Finally, we briefly touched upon alternatives to value-function estimation methods. We have identified a model-based approach as one possible alternative. The model-based approach, however, should be tailored so that the irrelevant aspects of the world are not paid attention to while learning the model. How to do this remains another very intriguing open problem.

Appendices

In the following appendices, we provide some auxiliary technical results that are omitted from the main body of the text. We start with a noncentral tail inequality (Appendix 7.A, Lemma 7.4), followed by a Bernstein-like concentration inequality for Hidden Markov Processes (Appendix 7.B, Theorem 7.6). We put these two results together to obtain a noncentral tail inequality for the considered class of dependent sequences (Appendix 7.C, Lemma 7.7). Finally in Appendix 7.D, we consider the problem of deriving high-probability excess-risk bounds in a regression setting.

7.A Noncentral Tail Inequalities

The following result shows that if a random variable X satisfies a Bernstein-like inequality, the probability distribution of X being ε-smaller than $(1-a)\mathbb{E}[X]$ or ε-larger than $(1+a)\mathbb{E}[X]$ (for $0 < a < 1$) decays with the rate $\exp(-c\,\varepsilon)$ for some c independent of ε. This should be contrasted with the slower $\exp(-c'\,\varepsilon^2)$ concentration rate of X around its expectation $\mathbb{E}[X]$ (for ε "small").

Lemma 7.4 (Noncentral Tail Inequality)**.** *Let X be a random variable whose expected value is nonnegative. Assume that for some $V > 0$ and for all $\varepsilon > 0$, X satisfies the following Bernstein-like tail inequality*

$$\mathbb{P}\{\mathbb{E}[X] - X \geq \varepsilon\} \leq \exp\left(-\frac{V\varepsilon^2}{\mathbb{E}[X] + \varepsilon}\right). \tag{7.13}$$

Then, for any $0 < a < 1$, $\varepsilon > 0$,

$$\mathbb{P}\left\{\mathbb{E}[X] - \frac{1}{1-a}X \geq \varepsilon\right\} \leq \exp\left(-\frac{V(1-a)a\varepsilon}{(1+a)}\right).$$

Similarly, if for some $V > 0$ and for all $\varepsilon > 0$ it holds that

$$\mathbb{P}\{X - \mathbb{E}[X] \geq \varepsilon\} \leq \exp\left(-\frac{V\varepsilon^2}{\mathbb{E}[X] + \varepsilon}\right) \tag{7.14}$$

then for all $0 < a < 1$ and $\varepsilon > 0$, it also holds that

$$\mathbb{P}\left\{\frac{1}{1+a}X - \mathbb{E}[X] \geq \varepsilon\right\} \leq \exp\left(-Va\varepsilon\right).$$

Proof. We have

$$\begin{aligned}
\mathbb{P}\left\{\mathbb{E}[X] - (1-a)^{-1}X \geq \varepsilon\right\} &= \mathbb{P}\left\{\mathbb{E}[X] - X \geq \varepsilon(1-a) + a\mathbb{E}[X]\right\} \\
&\leq \exp\left(-\frac{V\left((1-a)\varepsilon + a\mathbb{E}[X]\right)^2}{(1+a)\mathbb{E}[X] + (1-a)\varepsilon}\right) \\
&\leq \exp\left(-\frac{V\left((1-a)\varepsilon + a\mathbb{E}[X]\right)^2}{\left((1-a)\varepsilon + a\mathbb{E}[X]\right)\left(\frac{1+a}{a}\right)}\right) \\
&= \exp\left(-\frac{Va\left((1-a)\varepsilon + a\mathbb{E}[X]\right)}{1+a}\right) \\
&\leq \exp\left(-\frac{V(1-a)a\varepsilon}{1+a}\right),
\end{aligned}$$

where we used (7.13) to get the first inequality, added a positive value to upper bound the denominator in the second inequality, and used the fact that $\mathbb{E}[X] \geq 0$ to derive the last inequality.

Similarly, (7.14) leads to

$$\begin{aligned}
\mathbb{P}\left\{(1+a)^{-1}X - \mathbb{E}[X] > \varepsilon\right\} &= \mathbb{P}\left\{X - \mathbb{E}[X] > \varepsilon(1+a) + a\mathbb{E}[X]\right\} \\
&\leq \exp\left(-\frac{V\left((1+a)\varepsilon + a\mathbb{E}[X]\right)^2}{(1+a)\mathbb{E}[X] + (1+a)\varepsilon}\right) \\
&\leq \exp\left(-\frac{V\left((1+a)\varepsilon + a\mathbb{E}[X]\right)^2}{\left((1+a)\varepsilon + a\mathbb{E}[X]\right)\left(\frac{1+a}{a}\right)}\right) \\
&= \exp\left(-\frac{Va\left((1+a)\varepsilon + a\mathbb{E}[X]\right)}{1+a}\right) \\
&\leq \exp\left(-Va\varepsilon\right).
\end{aligned}$$

□

7.B Concentration Inequality for Hidden Markov Processes (HMPs)

The classical Bernstein inequality for independent and identically distributed sequences (e.g., Györfi et al. [2002, Appendix A]) can be shown to hold for the sequences of dependent random variables under various conditions. Such extensions are very useful when studying reinforcement learning algorithms when the standard assumption is that the data comes from some Markov chain. In this section we give such an extension based on Samson [2000].

Let $X_1, \ldots, X_n$ be a time-homogeneous Markov chain with transition kernel $P(\cdot|\cdot)$ taking values in some measurable space $\mathcal{X}$. We shall consider the concentration of the average of the Hidden-Markov Process

$$(X_1, f(X_1)), \ldots, (X_n, f(X_n)),$$

where $f : \mathcal{X} \to [0, B]$ is a fixed measurable function. To arrive at such an inequality, we need a characterization of how fast (X_i) forgets its past.

For $i > 0$, let $P^i(\cdot|x)$ be the i-step transition probability kernel: $P^i(A|x) = \mathbb{P}\{X_{i+1} \in A \,|\, X_1 = x\}$ (for all $A \subset \mathcal{X}$ measurable). Define the upper-triangular matrix $\Gamma_n = (\gamma_{ij}) \in \mathbb{R}^{n\times n}$ as follows:

$$\gamma_{ij}^2 = \sup_{(x,y)\in\mathcal{X}^2} \left\|P^{j-i}(\cdot|x) - P^{j-i}(\cdot|y)\right\|_{\mathrm{TV}}. \tag{7.15}$$

for $1 \leq i < j \leq n$ and let $\gamma_{ii} = 1$ $(1 \leq i \leq n)$.

Matrix Γ_n, and its operator norm $\|\Gamma_n\|$ w.r.t. the 2-norm, are measures of dependence for the random sequence $X_1, X_2, \ldots, X_n$. For example if the X_is are independent, $\Gamma_n = \mathbf{I}$ and $\|\Gamma_n\| = 1$. In general $\|\Gamma_n\|$, which appears in the forthcoming concentration inequalities for dependent sequences, can grow with n. Since the concentration bounds are homogeneous in $n/\|\Gamma_n\|^2$, a larger value $\|\Gamma_n\|^2$ means a smaller "effective" sample size. This motivates the following definition.

Definition 7.1. *We say that a time-homogeneous Markov chain* uniformly quickly forgets its past *if* $\tau = \sup_{n \geq 1} \|\Gamma_n\|^2 < +\infty$. *Further,* τ *is called the* forgetting time *of the chain.*

Conditions under which a Markov chain uniformly quickly forgets its past are of major interest. The following proposition, extracted from the discussion on pages 421–422 of the paper by Samson [2000], gives such a condition.

Proposition 7.5. *Let* μ *be some nonnegative measure on* $\mathcal{X}$ *with nonzero mass* μ_0. *Let* P^i *be the* i*-step transition kernel as defined above. Assume that there exists some integer* r *such that for all* $x \in \mathcal{X}$ *and all measurable sets* A,

$$P^r(A|x) \leq \mu(A). \tag{7.16}$$

Then,

$$\|\Gamma_n\| \leq \frac{\sqrt{2}}{1 - \rho^{\frac{1}{2r}}},$$

where $\rho = 1 - \mu_0$.

Meyn and Tweedie [2009] calls homogeneous Markov chains that satisfy the majorization condition (7.16) *uniformly ergodic*. We note in passing that there are other cases when $\sup_{n \geq 1} \|\Gamma_n\|$ is finite. Most notable, this holds when the Markov chain is contracting. The matrix Γ_n can also be defined for more general dependent processes and such that the theorem below remains valid. With such a definition, $\|\Gamma_n\|$ can be shown to be bounded for general Φ-dependent processes.

The following result is a trivial corollary of Theorem 2 of Samson [2000] (Theorem 2 is stated for empirical processes and can be considered as a generalization of Talagrand's inequality to dependent random variables):

Theorem 7.6. *Let* f *be a measurable function on* $\mathcal{X}$ *whose values lie in* $[0, B]$, $X_1, \ldots, X_n$ *be a homogeneous Markov chain taking values in* $\mathcal{X}$ *and let* Γ_n *be the matrix with elements defined by* (7.15). *Let*

$$Z = \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

Then, for every $\varepsilon \geq 0$,

$$\mathbb{P}\{Z - \mathbb{E}[Z] \geq \varepsilon\} \leq \exp\left(-\frac{\varepsilon^2 n}{2B\|\Gamma_n\|^2 (\mathbb{E}[Z] + \varepsilon)}\right),$$

$$\mathbb{P}\{\mathbb{E}[Z] - Z \geq \varepsilon\} \leq \exp\left(-\frac{\varepsilon^2 n}{2B\|\Gamma_n\|^2 \mathbb{E}[Z]}\right).$$

7.C Noncentral Tail Inequality for HMPs

By putting together the results of the last two sections we obtain the following noncentrail tail inequality for HMPs.

Lemma 7.7. *Let $X_1, X_2, \dots, X_n$ be a time-homogenous Markov chain taking values in some measurable space $\mathcal{X}$, and f be a measurable function with $0 \le f \le B$. Let $Z = \frac{1}{n}\sum_{i=1}^n f(X_i)$. Let Γ_n be the matrix with elements defined by (7.15). Then, for any $0 < a < 1$,*

$$\mathbb{P}\left\{\mathbb{E}[Z] - \frac{1}{1-a}Z \ge \varepsilon\right\} \le \exp\left(-\frac{(1-a)an\varepsilon}{2B\left\|\Gamma_n\right\|^2(1+a)}\right),$$
$$\mathbb{P}\left\{\frac{1}{1+a}Z - \mathbb{E}[Z] \ge \varepsilon\right\} \le \exp\left(-\frac{an\varepsilon}{2B\left\|\Gamma_n\right\|^2}\right).$$

Proof. According to Theorem 7.6,

$$\mathbb{P}\{Z - \mathbb{E}[Z] \ge \varepsilon\} \le \exp\left(-\frac{\varepsilon^2 n}{2B\left\|\Gamma_n\right\|^2(\mathbb{E}[Z] + \varepsilon)}\right)$$

and

$$\begin{aligned}\mathbb{P}\{\mathbb{E}[Z] - Z \ge \varepsilon\} &\le \exp\left(-\frac{\varepsilon^2 n}{2B\left\|\Gamma_n\right\|^2\mathbb{E}[Z]}\right) \\ &\le \exp\left(-\frac{\varepsilon^2 n}{2B\left\|\Gamma_n\right\|^2(\mathbb{E}[Z] + \varepsilon)}\right).\end{aligned}$$

These inequalities have the same form as the Bernstein-like inequality in Lemma 7.4 with the choice of $V = \frac{n}{2B\|\Gamma_n\|^2}$, and therefore imply the result. □

7.D Excess-Risk Estimation

Assumption A20 requires that BERMIN has access to a function $\bar{b}$ such that the excess risk $\|\tilde{Q}_k - T^*Q_k\|_\nu^2$ is below $\bar{b}(\delta)$ with probability at least $1-\delta$. In this section, we provide a general approach to come up with such a function. To avoid clutter, the notation of this section is not specialized to the reinforcement learning setup. The conversion, however, is straightforward: the function f^* here is the same as T^*Q_k $(k = 1, \dots, P)$ and the estimate $\hat{f}$ is the same as $\tilde{Q}_k$ that is returned by the REGRESS module in Algorithm 3. The random variables $X_i \in \mathcal{X}$ should be "read as" $(X_i, A_i) \in \mathcal{X} \times \mathcal{A}$ and $Y_i = \hat{T}^*Q_k(X_i, A_i)$.

The task of estimating the excess risk is difficult because what can directly be estimated based on the sample is the loss, and the expected loss of a predictor is larger than the excess risk by the loss of the best regressor, which is an unknown quantity. In this section we attack this problem under the assumption that the best regressor belongs to a known function space $\mathcal{F}$. We target the problem of simultaneously estimating a regressor and returning a high-probability risk bound for the excess risk of the computed regressor. If $\mathcal{F}$ was a "small" function space (e.g., it had a finite pseudo-dimension) then any procedure (such as empirical risk minimization) with known bounds on its estimation error would directly give a solution: The estimation error bound would provide a bound on the excess risk. To increase generality, here we consider the case when $\mathcal{F}$ is too large for such a simple approach to succeed, but $\mathcal{F}$ can be decomposed into an infinite sequence of "small" function spaces, $\mathcal{F}_k$: $\mathcal{F} = \cup_k \mathcal{F}_k$. Under this assumption the natural approach is to perform model selection and return the estimation error of the selected model. The reason this can be successful is because model selection will ultimately select a sufficiently complex model. We develop this idea in the rest of this section.

7.D.1 The Excess-Risk Estimation Algorithm

Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be a stationary, time-homogeneous Markov chain taking values in $\mathcal{X} \times [-B, B]$ for $\mathcal{X} \subset \mathbb{R}^d$ and let the regression function f^* be defined by $f^*(x) =$

Algorithm 5 REGRESS($\{\mathcal{D}_n, \mathcal{D}'_n\}, \{\mathcal{F}_1, \mathcal{F}_2, \dots\}, a_n, \tau, (C_k)$)

// Let $\{(X'_t, Y'_t)\}$ be the input-output pairs in $\mathcal{D}'_n$: $\mathcal{D}'_n = \{(X'_1, Y'_1), \dots, (X'_n, Y'_n)\}$.
for $k = 1, 2, \dots$ **do**
  $\hat{f}_k \leftarrow \mathsf{A}(\mathcal{D}_n, \mathcal{F}_k)$.
  $\bar{\mathcal{R}}_k = \frac{1}{(1-a_n)^2} \frac{1}{n} \sum_{i=1}^n (\hat{f}_k(X'_i) - Y'_i)^2$.
end for
$\hat{k} \leftarrow \mathrm{argmin}_{k \geq 1} \left[\bar{\mathcal{R}}_k + C_k\right]$.
Choose $\beta_1, \beta_2, \dots$ such that $\beta_k \geq 0$ and $\sum_{k \geq 1} \beta_k = 2/3$.
return $\hat{f}_{\hat{k}}$ and $\mathfrak{B}_{\hat{k}}(n, \cdot\, \beta_{\hat{k}}, \tau)$

$\mathbb{E}[Y_i | X_i = x]$. Let τ be an upper bound on the forgetting time of (X_i, Y_i) (cf. Appendix 7.B). Denote the stationary distribution underlying (X_i) by ν. Given $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$, the goal is to provide a good estimate $\hat{f}$ of f^* and a high confidence upper bound on the excess-risk

$$\|\hat{f} - f^*\|^2 \triangleq \|\hat{f} - f^*\|_{2,\nu}^2.$$

We assume that we are given a sequence of nested function spaces $(\mathcal{F}_k)$ and f^* is known to belong to their union $\cup_{k \geq 1} \mathcal{F}_k$. We further assume that we are given an algorithm A, which, given $\mathcal{F}_k$, δ, and a dataset of n points, returns an estimate $\hat{f}_k$ of f^* that belongs to $\mathcal{F}_k$. We further assume that for any $k \geq 1$ there exist functions $\mathfrak{A}_k$ and $\mathfrak{B}_k$ such that for any $0 < \delta \leq 1$,

$$L_k \triangleq \|\hat{f}_k - f^*\|^2 \leq \mathfrak{A}_k(f^*) + \mathfrak{B}_k(n, \delta, \tau) \tag{7.17}$$

holds with probability $1 - \delta$ and that the value $\mathfrak{B}_k(n, \delta, \tau)$, which possibly depends on the data, can be computed at any arguments (n, δ, τ) and hence is available to our algorithm. No similar assumption is made about function $\mathfrak{A}_k$.

The algorithm that we propose works with the data split in half: The first half, $\mathcal{D}_n$, is used to find the candidates $\hat{f}_k$ (by calling A), while the second half is used to run the model-selection algorithm to approximately select the candidate with the smallest excess risk. Finally, the algorithm returns the function $\mathfrak{B}_k(n, \cdot \beta_k, \tau)$ for the selected value of k as the high-probability bound on the excess-risk returns. Here, $\beta_k \geq 0$, $\sum_{k \geq 1} \beta_k = 2/3$ determines the *prior* distribution of the error probability δ. The algorithm is given as Algorithm 5. For simplicity, we assume that the full dataset, $\mathcal{D}_n \cup \mathcal{D}'_n$ holds $2n$ data points.

Bounds of the type (7.17) are of major interest in the theory of regression estimation. The first term, which depends only on k and f^* and is independent of n and δ corresponds to the so-called *approximation error* and shows how well one can approximate f^* with elements of $\mathcal{F}_k$. The second term is a bound on the error resulting from using a finite sample, i.e., it bounds the *estimation error*. When the sample is made of a sequence of independent, identically distributed random variables, there are many results in the literature that can provide bounds of the type (7.17), e.g., Györfi et al. 2002; van de Geer 2000; Lugosi and Wegkamp 2004; Bartlett et al. 2005. The case of dependent sample is much less explored. However, since at the heart of most result are exponential tail inequalities and most exponential tail inequalities available for the independent case have been extended to the dependent case, one expects that with some work existing bounds can be readily extended to the dependent case (see Chapter 4 for some recent results along this direction and a discussion of some prior work).

7.D.2 Theoretical Analysis of the Excess Error Estimator

The purpose of this section is to prove that under some technical conditions the regression estimate returned by Algorithm 5 satisfies an oracle-like property and the returned bound is a proper high-probability bound on the excess risk of the resulting estimator. The first

part of the statement follows easily from Theorem 7.1 and Lemma 7.7. The proof of this part is included mainly for the sake of completeness. The main novelty is the second part. The main idea underlying the proof of the second part is that for n large enough, with high probability $\hat{k}$ will be such that $\mathfrak{A}_{\hat{k}}(f^*) = 0$ and thus, by inequality (7.17), $\mathfrak{B}_{\hat{k}}(n, \delta\beta_{\hat{k}}, \tau)$ will bound the excess risk $L_{\hat{k}}$.
The assumptions under which we prove our result are as follows:

Assumption A22

Assumptions on the data:

1. $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$, $\mathcal{D}'_n = \{(X'_1, Y'_1), \dots, (X'_n, Y'_n)\}$, $X_i, X'_i \in \mathcal{X}$, $|Y_i|, |Y'_i| \le B$ for some $B > 0$.
2. $\mathcal{D}_n$ and $\mathcal{D}'_n$ are independent.
3. (X'_i, Y'_i) is a time-homogenous, stationary Markov chain and its forgetting time is upper bounded by τ. We denote by ν the stationary distribution underlying (X'_i) and we let $\|\cdot\| = \|\cdot\|_\nu$.

Assumptions on $(\mathcal{F}_k)$ and the regressor function f^*:

1. The function spaces $\mathcal{F}_1, \mathcal{F}_2, \dots$ hold measurable, real-valued functions with domain $\mathcal{X}$ bounded by $B > 0$.
2. The function $f^*(x) = \mathbb{E}\left[Y'_t | X'_t = x\right]$ belongs to $\cup_{k\ge 1}\mathcal{F}_k$.

Assumptions on algorithm A and functions $\mathfrak{A}_k$, $\mathfrak{B}_k$:

1. For any $n \ge 1$, $k \ge 1$, A returns a $\sigma(\mathcal{D}_n)$-measurable function $\hat{f}_k$ that belongs to $\mathcal{F}_k$ and the error bound (7.17) holds for this function with probability $1 - \delta$.
2. The functions $\mathfrak{A}_k$ are such that for some $C > 1$, $\mathfrak{A}_k(f^*) \le C \inf_{f\in\mathcal{F}_k} \|f - f^*\|^2$ holds for all $k \ge 1$ and $\mathfrak{A}_k(\cdot) \ge \mathfrak{A}_{k+1}(\cdot)$ holds for any $k \ge 1$.
3. The function $\mathfrak{B}_k(n, \delta, \tau) \xrightarrow{n\to\infty} 0$ is a decreasing function of n and an increasing function of τ.

Note that we did not need to assume that the function spaces are nested, because in the proof all we need is that the functions $(\mathfrak{A}_k)$ satisfy $\mathfrak{A}_{k+1} \le \mathfrak{A}_k$. If $\mathfrak{A}_k(f^*) = C \inf_{f\in\mathcal{F}_k} \|f - f^*\|^2$, then the nestedness of $(\mathcal{F}_k)$ implies that $(\mathfrak{A}_k)$ is a pointwise decreasing sequence of functions.

The following theorem is the main result of this section.

Theorem 7.8. *Assume that the conditions listed in Assumption A22 hold and the value of a_n given to the algorithm depends on n and in particular $a_n = cn^{-1/2}$ with some $c > 0$. Assume that the penalty factors, $C_k = C_k(n)$, passed to Algorithm 5 are such that for any fixed k, $C_k(n)$ is a strictly decreasing function of n and for any fixed n,*

$$S_n = \sum_{k\ge 1} \exp\left(-\frac{(1-a_n)^2 a_n n}{8B^2(1+a_n)\tau} C_k(n)\right) < \infty. \tag{7.18}$$

Let $\hat{f}$ and $\hat{b}$ be the pair returned by Algorithm 5. Then, the following hold:
(A) For any $0 < \delta \le 1$,

$$\left\|\hat{f} - f^*\right\|^2 \le (1-a_n^2) \min_{k\ge 1}\left[\frac{\left\|\hat{f}_k - f^*\right\|^2}{(1-a_n)^2} + 2C_k(n)\right] + \frac{2a_n}{1-a_n} L(f^*) + \frac{16B^2(1+a_n)\tau \ln(\frac{2S_n}{\delta})}{(1-a_n)a_n n}$$

holds with probability at least $1 - \delta$, where $L(f) = \mathbb{E}\left[(f(X'_1) - Y'_1)^2\right]$.
(B) Fix $0 < \delta \le 1$. Then, there exists $n_0 = n_0(f^*, \delta) \ge 1$ such that for any $n \ge n_0$, the inequality $\left\|\hat{f} - f^*\right\|^2 \le \hat{b}(\delta)$ holds with probability at least $1 - \delta$.

Note that by selecting $a_n \propto n^{-1/2}$, Part (A) shows that the procedure's excess error above the oracle's performance is $O(n^{-1/2})$.

Proof. Let $\hat{k}$ be the index selected by Algorithm 5. A standard calculation shows that $\mathbb{E}\left[\bar{\mathcal{R}}_k | \mathcal{D}_n\right] = \frac{1}{(1-a_n)^2} L(\hat{f}_k)$, where for any fixed function f, $L(f) = \mathbb{E}\left[(f(X_1') - Y_1')^2\right]$ denotes the squared prediction loss of f. Our goal is to apply Theorem 7.1 to derive a bound on $L(f_{\hat{k}})$ and then relate $L(f_{\hat{k}})$ to the excess risk $L_{\hat{k}}$. We verify the conditions of Theorem 7.1. As before, the theorem is applied to the probability space obtained by conditioning w.r.t. $\mathcal{D}_n$. Let us first verify conditions (7.2)-(7.3) of Theorem 7.1, which connect $L(\hat{f}_k)$ and $\mathcal{R}_k$. In order to verify these conditions, we use Lemma 7.7. Let $g : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ be defined by $x \mapsto (\hat{f}_k(x) - y)^2$. By assumption, the range of g is a subset of $[0, 4B^2]$. Hence, applying Lemma 7.7 to $Z = 1/n \sum_{i=1}^n g(X_i', Y_i')$, exploiting that $(1-a_n)^2 \bar{\mathcal{R}}_k = Z$, after some algebra we get that for all $\varepsilon > 0$, the following inequalities are satisfied:

$$\mathbb{P}\left\{ L(\hat{f}_k) - (1-a_n)\bar{\mathcal{R}}_k > \varepsilon \middle| \mathcal{D}_n \right\} \le \exp\left(-\frac{(1-a_n)a_n}{8B^2\tau(1+a_n)} \varepsilon \right),$$
$$\mathbb{P}\left\{ \frac{1}{1+a_n}\bar{\mathcal{R}}_k - \mathbb{E}\left[\bar{\mathcal{R}}_k | \mathcal{D}_n\right] > \varepsilon \middle| \mathcal{D}_n \right\} \le \exp\left(-\frac{(1-a_n)^2 a_n n}{8B^2\tau} \varepsilon \right).$$

Choosing $c_1, c_3 = 1$, $c_2 = \frac{(1-a_n)a_n n}{8B^2\tau(1+a_n)}$, and $c_4 = \frac{(1-a_n)^2 a_n n}{8B^2\tau}$, we see that conditions (7.2) and (7.3) of Theorem 7.1 are satisfied. Further, let c_5 (c_6) of Theorem 7.1 be defined as in (7.4) (respectively, as in (7.5)). Then, if $(C_k(n))$ is chosen such that (7.18) is satisfied, we also have $c_6 \le c_5 = S_n < +\infty$, as required. Therefore, Part (B) of Theorem 7.1 with the choice of $\alpha = \alpha(n, a_n, \delta)$, where

$$\alpha(n, a_n, \delta) = \frac{16B^2(1+a_n)\tau \ln(\frac{2S_n}{\delta})}{(1-a_n)a_n n}$$

implies that with probability $1-\delta$,

$$L(\hat{f}_{\hat{k}}) \le (1-a_n^2) \min_{k \ge 1} \left[\frac{1}{(1-a_n)^2} L(\hat{f}_k) + 2C_k(n) \right] + \alpha(n, a_n, \delta).$$

Subtract $L(f^*)$ from both sides and use that $L_k = L(\hat{f}_k) - L(f^*)$ to get

$$L_{\hat{k}} \le (1-a_n^2) \min_{k \ge 1} \left[\frac{1}{(1-a_n)^2} L_k + 2C_k(n) \right] + \frac{2a_n}{1-a_n} L(f^*) + \alpha(n, a_n, \delta).$$

This finishes the proof of Part (A).

Let us now prove Part (B). Fix some $0 < \delta \le 1$. Let $\mathcal{E}_1$ be the error event where

$$\left\| \hat{f}_{\hat{k}} - f^* \right\|^2 \le (1-a_n^2) \min_{k \ge 1} \left[\frac{\left\| \hat{f}_k - f^* \right\|^2}{(1-a_n)^2} + 2C_k(n) \right] + \frac{2a_n}{1-a_n} L(f^*) + \alpha(n, a_n, \delta/3) \quad (7.19)$$

fails to hold. By Part (A), $\mathbb{P}\{\mathcal{E}_1\} \le \delta/3$. Let $\mathcal{E}_2$ be the error event where one of the inequalities

$$\left\| \hat{f}_k - f^* \right\|^2 \le \mathfrak{A}_k(f^*) + \mathfrak{B}_k(n, \beta_k \delta, \tau), \quad k = 1, 2, \dots \quad (7.20)$$

fails to hold. By assumption and the choice of (β_k), $\mathbb{P}\{\mathcal{E}_2\} \le 2\delta/3$. Our goal is to show that for n large enough, outside of $\mathcal{E} = \mathcal{E}_1 \cup \mathcal{E}_2$, $\mathfrak{A}_{\hat{k}}(f^*) = 0$. Indeed, if this holds then outside of $\mathcal{E}$, $\left\| \hat{f}_{\hat{k}} - f^* \right\|^2 \le \mathfrak{A}_{\hat{k}}(f^*) + \mathfrak{B}_{\hat{k}}(n, \beta_{\hat{k}}\delta, \tau) = \mathfrak{B}_{\hat{k}}(n, \beta_{\hat{k}}\delta, \tau)$, which implies the desired statement.

In the rest of the proof, all of our derivations will be done on the event $\mathcal{E}^c$. Let k^* be the first index where $\mathfrak{A}_k(f^*) = 0$. Note that k^* is well-defined by our assumption that relates

$\mathfrak{A}_k(f^*)$ to the approximation errors, $\inf_{f\in\mathcal{F}_k}\|f-f^*\|^2$, and because $f^*\in\cup_{k\geq 1}\mathcal{F}_k$. If $k^*=1$, then $\hat{k}\geq k^*$ and thus $\mathfrak{A}_{\hat{k}}(f^*)=0$ holds, too. Therefore, from now on assume that $k^*>1$. From (7.19), it follows that

$$\left\|\hat{f}_{\hat{k}}-f^*\right\|^2 \leq (1-a_n^2)\left[\frac{\left\|\hat{f}_{k^*}-f^*\right\|^2}{(1-a_n)^2}+2C_{k^*}(n)\right]+\frac{2a_n}{1-a_n}L(f^*)+\alpha(n,a_n,\delta/3)\,.$$

By (7.20), we also have $\|\hat{f}_{k^*}-f^*\|^2\leq\mathfrak{A}_{k^*}(f^*)+\mathfrak{B}_{k^*}(n,\beta_{k^*}\delta,\tau)=\mathfrak{B}_{k^*}(n,\beta_{k^*}\delta,\tau)$. Chaining these inequalities gives

$$\left\|\hat{f}_{\hat{k}}-f^*\right\|^2 \leq (1-a_n^2)\left[\frac{\mathfrak{B}_{k^*}(n,\beta_{k^*}\delta,\tau)}{(1-a_n)^2}+2C_{k^*}(n)\right]+\frac{2a_n}{1-a_n}L(f^*)+\alpha(n,a_n,\tfrac{\delta}{3}). \quad (7.21)$$

Let n_0 be the first integer such that the right-hand side of (7.21) is strictly below $0<\mathfrak{A}_{k^*-1}(f^*)/C$. Such an index exists because the right-hand side of (7.21) converges to zero as $n\to\infty$. Since $\hat{f}_{\hat{k}}\in\mathcal{F}_{\hat{k}}$, we have $\inf_{f\in\mathcal{F}_{\hat{k}}}\|f-f^*\|^2\leq\|\hat{f}_{\hat{k}}-f^*\|^2$. Therefore, if $n\geq n_0$, $\hat{k}=\hat{k}_n$ is such that $\mathfrak{A}_{\hat{k}}(f^*)\leq C\inf_{f\in\mathcal{F}_{\hat{k}}}\|f-f^*\|^2\leq C\|\hat{f}_{\hat{k}}-f^*\|^2<\mathfrak{A}_{k^*-1}(f^*)$ and thus, by the definition of k^*, $\mathfrak{A}_{\hat{k}}(f^*)=0$, thus finishing the proof. □

Chapter 8

Concluding Remarks

In this thesis, we have investigated a regularization-based approach to solve RL/Planning problems with large state spaces. We developed a regularized AVI algorithm, RFQI, in Chapter 5 and two regularized API algorithms, namely REG-BRM and REG-LSPI, in Chapter 6. All these algorithms are formulated as regularized optimization problems with a rather general choice of the function space and regularizer. When the function space was an RKHS, we provided closed-form solutions for the corresponding optimization problems. The main emphasis of this thesis has been on the analysis of the statistical properties of the proposed algorithms. Under generic assumptions on the capacity of the function space and some properties of the MDP, we provided the performance loss upper bounds for RFQI and REG-LSPI. We also investigated the model selection problem for RL/Planning problems in Chapter 7. We developed and studied a complexity-regularization algorithm to find the minimum of the Bellman error among a set of candidate action-value functions.

In our endeavor, we also addressed some other issues that are not specific to regularization-based RL/Planning algorithms but have been relevant to our discussion:

- We examined how the errors at each iteration of AVI/API would affect the quality of the resulting policy. These results can be used for any AVI/API algorithm – parametric or nonparametric (Chapter 3).
- We studied regularized least-squares regression with the β-mixing input data (Chapter 4). We provided an error upper bound on the excess error and showed that under certain assumptions, most importantly having an exponentially fast mixing process, the convergence rate is asymptotically the same as the rate for the i.i.d. input process. This result is used to analyze the RFQI algorithm.

From a philosophical standpoint, this work provided a concrete formalism of the **Occam's Razor** principle: Given some data from interaction with an MDP, the regularization-based algorithms prefer "simpler" explanations for the estimate of the value function. This work extends the previous formalisms, which have focused on the prediction problems in the supervised learning setting, to the new ground of sequential decision-making problems and control. We have seen that the concept of simplicity is context-dependent as it is specified by the choice of the function space and the corresponding regularizer. Varying these two leads to different ways of preferring simpler solutions.

8.1 Suggestions for Future Research

This thesis opens up several possibilities for further investigations. We have already commented on some of them in the concluding sections of each chapter, but in order to make them more accessible, we briefly review them here too.

Computational Efficiency

The focus of this work has been on laying the foundation of regularization-based RL/Planning algorithms and studying their statistical properties. Although our algorithms are computationally tractable, their naive implementations may not lead to a computer program that can handle millions of data samples with the contemporary hardware technology. Therefore, one has to address this issue by devising elegant numerical algorithms. Three possible approaches are 1) to use iterative algorithms in conjunction with fast matrix-vector multiplication (e.g., using ideas from the Fast Multipole Methods), 2) to sparsify the samples and work with a representative subset of samples, and 3) to apply stochastic gradient-based algorithms. See Section 5.6 for more detail.

Continuous Action Space

Many real-world applications, including almost all control engineering and robotics problems, are best described by a continuous action space. Discretizing a continuous action space and using an algorithm that is designed to work with finite action spaces does not scale well with the dimensionality of the action space.

Our general formulation of RFQI, REG-LSPI, and REG-BRM allows us to choose a function space, such as an RKHS, that has a continuous action domain. The difficulty, however, is that the current analysis assumes that one may find $\max_{a' \in \mathcal{A}} Q(x, a)$ for a given action-value function Q. Of course finding the maximizer is not feasible for a large $\mathcal{A}$ and an arbitrary Q function – for this is an instance of global optimization problem. One possibility is to search locally around some given action a_0 and return the value of the local maximizer. This inexact policy improvement requires new error propagation results.

Partial Observable Problems

The working assumption of this thesis has been that the state of the system is accessible. In many real-world problems, however, this assumption does not hold. What we have instead is an observation that might not be a sufficient statistic of the history. Showing that how the current algorithms can be extended, if needed, to handle these types of problems is the topic of future research.

Online Regularized RL/Planning Algorithms

Our methods are stated and most of our results are proven in the offline learning scenario. One can argue that many real-world problems are better described by an agent in a continual interaction with its environment.

To have an online RL/Planning algorithms, one should modify our algorithms so that they incrementally update their action-value function estimates in a computationally inexpensive manner. One difficulty for nonparametric methods such as ours is that the effective function space is changing when new samples are added. Efficient re-estimation of the value function in this sequence of ever-changing function spaces deserves further investigation.

Other Regularizers

The formulations of RFQI, REG-LPSI, and REG-BRM are general and allow different choices of the function space and regularizers. Moreover, from the statistical point of view our results hold for a large class of function spaces and regularizers that satisfy the specified capacity condition and a few other assumptions. Nonetheless, one should ask two questions before using a function space and a regularizer:

1. Does this pair of the function space and the regularizer provide a natural way of controlling the complexity for the given problem?

2. How can the corresponding optimization problems be solved?

For instance, consider the choice of the l_1-regularization. The first question asks whether sparsity-inducing regularization is an appropriate choice for the RL/Planning problem in hand. If the basis functions are an over-complete dictionary, such as wavelet basis, the choice of the l_1-regularization seems viable. Or if we know that many input variables are irrelevant to the representation of the action-value function, this choice is also appropriate. The second question asks if we can efficiently solve the optimization problems with the l_1-regularizer. For RFQI, this is not a major issue as we essentially should solve a LASSO problem. This is not, however, the case for REG-LSPI and REG-BRM. See the discussion of Section 6.5.

Model-based Policy Selection Algorithm

The model selection approach we have advocated in Chapter 7 is an indirect one because it selects the best policy according to the estimate of its Bellman error. As briefly suggested in Section 7.5, a more direct approach is possible too: Learn the model of the environment, generate virtual samples from the learned model, and use these samples to assess the quality of different policies. If the learned model is a close approximation to the environment, which calls for an appropriate model selection itself, one might expect to get an accurate evaluation of the performance for each policy. Therefore, the ranking of the performance based on the learned model would be close to the true ranking of the policies. A rigorous study of this model-based policy selection algorithm is required for better understanding of the relative advantages and disadvantages of these two approaches.

Regularities of MDP

In this thesis, we related the sample complexity of learning to the properties of the action-value function space (such as its metric entropy) and the norm-expansion property of the Bellman operator in that space (Assumptions A11 and A19) – see Theorems 5.8 and 6.6. An interesting question is how these properties are related to the regularities of the transition probability kernel P and the reward kernel $\mathcal{R}$. We partially addressed this question for a certain class of MDPs, which we called convolutional MDPs (Section 6.G), but more studies is required.

Aside theoretical appeal, an answer to this question might also have practical implications. It helps us to better understand what types of function spaces should we expect to confront in RL/Planning problems. This can be used to design a better set of candidate models (cf. Remark 7.3). For instance if we know that the reward function is quadratic in the magnitude of the state error and the action (control signal) and the MDP is a linear system with additive Gaussian noise (e.g., classical Linear Quadratic Regulator (LQR) setup), what is the right choice of the RKHS for this problem and what are the values of L_P and L_R in Assumptions A11 and A19?[1]

Another related open question is to characterize the dynamics of concentrability coefficients defined in Chapter 3 and to relate it to the properties of the transition probability kernel P.

Lower Bounds for RL/Planning

The focus of this thesis has been on providing sample complexity upper bounds on the performance loss. To show the tightness of upper bounds, one should also provide a matching lower bound. This problem can be seen from two aspects. The first is whether the value estimation task in an RL/Planning problem is done optimally or not, and the other is

[1] We know that the optimal cost-to-go function of the LQR problem is quadratic in state.

whether the optimal estimation of the value function is necessary to perform optimally (i.e., the optimal convergence rate for the performance loss).

The value estimation task in an RL/Planning problem can be considered as a generalization of regression problems (a regression problem can be reduced to a value estimation problem for the corresponding RL/Planning problem that has $\gamma = 0$). Our upper bounds show that the behavior of the sample complexity is essentially the same as those suggested by the usual lower bounds in the regression literature. This indicates that the estimation part of our results are optimal when both $\mathcal{R}$ and P are unknown. Nevertheless, one may consider a different scenario when P is not known *but* the reward function r is known a priori. In this scenario, whenever $\gamma = 0$, there is nothing to learn: we already know the best estimate of the action-value function and it is $Q(x, a) = r(x, a)$. But when $\gamma > 0$, we still require to estimate the action-value function based on samples. Thus, the question is what can be said about the lower bound on the sample complexity for the value estimation task?

Nevertheless, it is not clear whether the optimal estimation of the value function is necessary for the optimal performance of the agent. One can easily imagine situations where the value function is estimated inaccurately, but the selected policy (e.g., the greedy policy w.r.t the estimated action-value function) is the optimal one. Therefore, one can rightfully ask what the true sample complexity of an RL/Planning problem is regardless of whether a value-based approach is followed or not.

Exploration-Exploitation Tradeoff

Throughout the thesis, we assumed that the data sampling distribution ν is fixed. As it is apparent from the definition of concentrability coefficients in Chapter 3, the choice of ν has direct effect on the quality of the resulting policy.

The choice of ν can be studied in two different scenarios. In the first scenario, the goal is to find the best policy given a finite budget of actively collected samples. This is closer to the spirit of offline learning. The goal of the second scenario is to minimize the regret in online learning. This scenario is usually called the exploration-exploitation tradeoff problem. Efficient solutions for both of these problems are crucial for an agent that solves large RL/Planning problems.

Regularized RL algorithms for Average Reward MDPs and SMDPs

Many sequential decision-making problems are best described not by discounted MDPs, but by other classes of problems such as *average reward* MDPs or *Semi-Markov Decision Processes (SMDP)*. Extending our current methods to these classes of sequential decision-making problems is the topic of future research.

Bibliography

Anestis Antoniadis. Wavelet methods in statistics: Some recent developments and their applications. *Statistical Surveys*, 1:16 – 55, 2007. 66

András Antos, László Györfi, and Michael Kohler. Lower bounds on the rate of convergence of nonparametric regression estimates. *Journal of Statistical Planning and Inference*, 83 (1):91 – 100, 2000. 160

András Antos, Csaba Szepesvári, and Rémi Munos. Value-iteration based fitted policy iteration: Learning with a single trajectory. In *IEEE Symposium on Approximate Dynamic Programming and Reinforcement Learning (ADPRL)*, pages 330–337. IEEE, 2007. 117

András Antos, Rémi Munos, and Csaba Szepesvári. Fitted Q-iteration in continuous action-space MDPs. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems (NIPS - 20)*, pages 9–16, Cambridge, MA, 2008a. MIT Press. 15, 20, 68

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 Journal*, 71:89–129, 2008b. 16, 20, 21, 23, 24, 26, 33, 34, 35, 36, 74, 75, 76, 77, 82, 86, 88, 89, 117, 132

Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2009. 56, 69, 126

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337 – 404, May 1950. 164

Jean-Yves Audibert and Alexander B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007. 162

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 89–96, 2009. 17

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. 79, 88, 89

Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 30–37. Morgan Kaufmann, 1995. 74

Andrew R. Barron. Complexity regularization with application to artificial neural networks. In G. Roussas, editor, *Nonparametric Function Estimation and Related Topics*, pages 561–576. Kluwer Academic Publishers, 1991. 116, 124, 129

Andrew R. Barron, Cong Huang, Jonathan Q. Li, and Xi Luo. The MDL principle, maximum likelihoods, and statistical risk. In Peter Grünwald, Petri Myllymäki, Ioan Tabus, Marcelo Weinberger, and Bin Yu, editors, *Festschrift in Honor of Jorma Rissanen on the Occasion of his 75th Birthday*, TICSP Series #38. Tampere International Center for Signal Processing, 2008. 129

Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002. 158

Peter L. Bartlett and Ambuj Tewari. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence*, 2009. 17

Peter L. Bartlett, Stéphane Boucheron, and Gábor Lugosi. Model selection and error estimation. *Machine Learning*, 48(1-3):85–113, 2002. 56, 116, 123, 124

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005. 138

Jonathan Baxter and Peter L. Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, pages 319–350, 2001. 13

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. 68, 91

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In Bernhard Schölkopf, John C. Platt, Thomas Hoffman, Bernhard Schölkopf, John C. Platt, and Thomas Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 137–144. MIT Press, 2006. 74

Sergei Natanovich Bernstein. Sur l'extension du théorème limite du calcul des probabilités aux sommes de quantités dépendantes. *Mathematische Annalen*, 97:1–59, 1927. 38

Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 2, chapter Approximate Dynamic Programming. Athena Scientific, 3rd edition, May 2010. 7, 17

Dimitri P. Bertsekas and Steven E. Shreve. *Stochastic Optimal Control: The Discrete-Time Case*. Academic Press, 1978. 7, 8, 9, 10, 11, 15, 16

Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming (Optimization and Neural Computation Series, 3)*. Athena Scientific, 1996. 7, 15, 23, 33

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. 3, 18, 118, 158

Leon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 161–168. MIT Press, Cambridge, MA, 2008. 68, 91

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005. 162

Richard C. Bradley. Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys*, 2:107–44, 2005. 37

Steven J. Bradtke and Andrew G. Barto. Linear least-squares algorithms for temporal difference learning. *Machine Learning*, 22:33–57, 1996. 16, 76

Jeffrey B. Burl. *Linear Optimal Control: H_2 and H_∞ Methods*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1998. 15

Lucian Buşoniu, Robert Babuška, Bart De Schutter, and Damien Ernst. *Reinforcement Learning and Dynamic Programming Using Function Approximators*. CRC Press, 2010a. 7, 17

Lucian Buşoniu, Damien Ernst, Bart De Schutter, and Robert Babuska. Approximate dynamic programming with a fuzzy parameterization. *Automatica*, 46(5):804 – 814, 2010b. 19

François Chaumette and Seth Hutchinson. Visual servoing and visual tracking. In Siciliano and Khatib [2008], pages 563–583. 1

Fan R. K. Chung. *Spectral Graph Theory (CBMS Regional Conference Series in Mathematics, No. 92)*. American Mathematical Society, 1997. 19

Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1 – 49, 2002. 164

Sanjoy Dasgupta and Yaov Freund. Random projection trees and low dimensional manifolds. In *Proceedings of ACM Symposium on Theory of Computing (STOC)*, pages 537–546. ACM, 2008. 162

Daniela Pucci de Farias and Benjamin Van Roy. On the existence of fixed points for approximate value iteration and temporal-difference learning. *Journal of Optimization Theory and Application*, 105(3):589–608, 2000. 20

Daniela Pucci de Farias and Benjamin Van Roy. The linear programming approach to approximate dynamic programming. *Operations Research*, 51(6):850–865, 2003. 15

Ronald A. Devore. Nonlinear approximation. *Acta Numerica*, 7:51–150, 1998. 15, 17, 163

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Applications of Mathematics: Stochastic Modelling and Applied Probability. Springer-Verlag New York, 1996. 158

David L. Donoho and Iain M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200 – 1224, December 1995. 66, 67, 161

David L. Donoho and Iain M. Johnstone. Minimax estimation via wavelet shrinkage. *The Annals of Statistics*, 26(3):879 – 921, 1998. 66

Paul Doukhan. *Mixing: Properties and Examples*, volume 85 of *Lecture Notes in Statistics*. Springer-Verlag, Berlin, 1994. 37, 82, 92

Bradley Efron, Trevor Hastie, Iain M. Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407 – 499, 2004. 90

Yaakov Engel, Shie Mannor, and Ron Meir. Reinforcement learning with Gaussian processes. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pages 201–208. ACM, 2005. 19, 68, 91, 115

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6:503–556, 2005. 15, 19, 53

Eyal Even-Dar and Yishay Mansour. Learning rates for Q-learning. *Journal of Machine Learning Research*, 5:1–25, 2003. 20

Amir-massoud Farahmand. Action-gap phenomenon in reinforcement learning. In J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems (NIPS - 24)*, pages 172–180. Curran Associates, Inc., 2011. 88

Amir-massoud Farahmand and Csaba Szepesvári. Model selection in reinforcement learning. *Machine Learning Journal*, 85(3):299–332, 2011. 5

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. URL http://dx.doi.org/10.1016/j.jspi.2011.08.007. 5, 52

Amir-massoud Farahmand, Azad Shademan, and Martin Jägersand. Global visual-motor estimation for uncalibrated visual servoing. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1969–1974. IEEE, 2007a. 1

Amir-massoud Farahmand, Csaba Szepesvári, and Jean-Yves Audibert. Manifold-adaptive dimension estimation. In *ICML '07: Proceedings of the 24th international conference on Machine learning*, pages 265–272, New York, NY, USA, 2007b. ACM. 162

Amir-massoud Farahmand, Csaba Szepesvári, and Jean-Yves Audibert. Toward manifold-adaptive learning. In *NIPS Workshop on Topology learning*, Whistler, Canada, December 2007c. 162

Amir-massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. Regularized fitted Q-iteration: Application to planning. In Sertan Girgin, Manuel Loth, Rémi Munos, Philippe Preux, and Daniil Ryabko, editors, *Recent Advances in Reinforcement Learning, 8th European Workshop, EWRL 2008*, volume 5323 of *Lecture Notes in Computer Science*, pages 55–68. Springer, 2008. 5, 19, 68, 90

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. 5, 19, 90

Amir-massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. Regularized policy iteration. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems (NIPS - 21)*, pages 441–448. MIT Press, 2009b. 5, 19, 20, 90

Amir-massoud Farahmand, Azad Shademan, Martin Jägersand, and Csaba Szepesvári. Model-based and model-free reinforcement learning for visual servoing. In *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, pages 2917–2924, May 2009c. 1, 5, 68

Amir-massoud Farahmand, Rémi Munos, and Csaba Szepesvári. Error propagation for approximate policy and value iteration. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems (NIPS - 23)*, pages 568–576. 2010. 4, 32, 74, 75

Amir-massoud Farahmand, Csaba Szepesvári, and Jean-Yves Audibert. Nearest neighborhood methods for manifold-adaptive dimension estimation and regression. Under Preparation, 2011. 162

Matthieu Geist and Bruno Scherrer. ℓ_1-penalized projected Bellman residual. In Scott Sanner and Marcus Hutter, editors, *Recent Advances in Reinforcement Learning*, volume 7188 of *Lecture Notes in Computer Science*, pages 89–101. Springer Berlin Heidelberg, 2012. 79

Alborz Geramifard, Michael Bowling, Michael Zinkevich, and Richard S. Sutton. iLSTD: Eligibility traces and convergence analysis. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 441–448. MIT Press, Cambridge, MA, 2007. 16

Mohammad Ghavamzadeh and Yaakov Engel. Bayesian actor-critic algorithms. In Zoubin Ghahramani, editor, *ICML '07: Proceedings of the 24th Annual International Conference on Machine Learning*, pages 297–304. Omnipress, 2007a. 13

Mohammad Ghavamzadeh and Yaakov Engel. Bayesian policy gradient algorithms. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 457–464. MIT Press, Cambridge, MA, 2007b. 13

Mohammad Ghavamzadeh, Alessandro Lazaric, Rémi Munos, and Matthew Hoffman. Finite-sample analysis of lasso-TD. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning (ICML)*, ICML '11, pages 1177–1184, New York, NY, USA, June 2011. ACM. ISBN 978-1-4503-0619-5. 79, 88, 89, 90, 91

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Verlag, New York, 2002. 3, 4, 15, 18, 36, 41, 42, 43, 44, 45, 50, 51, 58, 75, 84, 92, 94, 118, 121, 135, 138, 158, 159, 161, 164, 165

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2001. 3, 4, 18, 118, 158

Meisner Heidrich-Meisner and Christian Igel. Hoeffding and Bernstein races for selecting policies in evolutionary direct policy search. In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pages 401–408, New York, NY, USA, 2009. ACM. 13

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 1970. 19, 55, 79

Vivian Hutson, John Sydney Pym, and Michael J. Cloud. *Applications of Functional Analysis and Operator Theory (Second Edition)*. Elsevier, 2005. 166

Tommi Jaakkola, Michael I. Jordan, and Satinder P. Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neuralal Computation*, 6:1185–1201, November 1994. ISSN 0899-7667. 20

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563 – 1600, 2010. 17

Jeff Johns, Christopher Painter-Wakefield, and Ronald Parr. Linear complementarity for regularized policy evaluation and improvement. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems (NIPS - 23)*, pages 1009–1017. 2010. 91

Tobias Jung and Daniel Polani. Least squares SVM for least squares TD learning. In *Proceedings of the 17th European Conference on Artificial Intelligence*, pages 499–503, 2006. 16, 19, 68, 90, 91, 115

Sham Kakade. A natural policy gradient. In *NIPS*, pages 1531–1538, 2001. 13

Philipp W. Keller, Shie Mannor, and Doina Precup. Automatic basis function construction for approximate dynamic programming and reinforcement learning. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 449–456, New York, NY, USA, 2006. ACM. 117

Charles C. Kemp, Paul Fitzpatrick, Hirohisa Hirukawa, Kazuhito Yokoi, Kensuke Harada, and Yoshio Matsumoto. Humanoids. In Siciliano and Khatib [2008], pages 1307–1333. 1

Hassan K. Khalil. *Nonlinear Systems (3rd Edition)*. Prentice Hall, 2001. 91

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. 38, 41, 42, 43, 50

Michael Kohler, Adam Krzyżżk, and Dominik Schäfer. Application of structural risk minimization to multivariate smoothing spline regression estimates. *Bernoulli*, 8(4):475–489, 2002. 3, 45, 49

Vladimir Koltchinskii. 2004 IMS medallion lecture: Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34:2593–2656, 2006. 158

J. Zico Kolter and Andrew Y. Ng. Regularization and feature selection in least-squares temporal difference learning. In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pages 521–528. ACM, 2009. 16, 19, 79, 90, 115

Vijay R. Konda and John N. Tsitsiklis. On actor-critic algorithms. *SIAM Journal on Control and Optimization*, pages 1143–1166, 2001. 13

John Lafferty and Larry Wasserman. Challenges in statistical machine learning. *Statistica Sinica*, 16:307–322, 2006. 158, 161

Michail G. Lagoudakis and Ronald Parr. Least-squares policy iteration. *Journal of Machine Learning Research*, 4:1107–1149, 2003. 16, 17, 34, 75, 76, 116, 117, 122

Sascha Lange and Martin Riedmiller. Deep auto-encoder neural networks in reinforcement learning. In *International Joint Conference on Neural Networks (IJCNN 2010)*, 2010. 19

Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Finite-sample analysis of least-squares policy iteration. *Journal of Machine Learning Research*, 13:3041–3074, October 2012. 88, 89

Wee Sun Lee, Peter L. Bartlett, and Robert C. Williamson. The importance of convexity in learning with squared loss. *IEEE Transactions on Information Theory*, 44(5):1974–1980, 1998. 160

Wee Sun Lee, Peter L. Bartlett, and Robert C. Williamson. Correction to *the importance of convexity in learning with squared loss. IEEE Transactions on Information Theory*, 54 (9):4395, 2008. 160

Yuxi Li, Csaba Szepesvári, and Dale Schuurmans. Learning exercise policies for American options. In *International Conference on Artificial Intelligence and Statistics (AISTATS-09)*, 2009. 2

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. 19, 90, 115

Gábor Lugosi and Marten Wegkamp. Complexity regularization via localized random penalties. *The Annals of Statistics*, 32:1679–1697, 2004. 116, 124, 138

Hamid Reza Maei, Csaba Szepesvári, Shalabh Bhatnagar, Doina Precup, David Silver, and Richard S. Sutton. Convergent temporal-difference learning with arbitrary smooth function approximation. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems (NIPS - 22)*, pages 1204–1212. 2009. 20

Hamid Reza Maei, Csaba Szepesvári, Shalabh Bhatnagar, and Richard S. Sutton. Toward off-policy learning control with function approximation. In Johannes Fürnkranz and Thorsten Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pages 719–726, Haifa, Israel, June 2010. Omnipress. 20

Sridhar Mahadevan and Mauro Maggioni. Proto-value functions: A Laplacian framework for learning representation and control in Markov decision processes. *Journal of Machine Learning Research*, 8:2169–2231, 2007. 16, 18, 19

Odalric Maillard, Rémi Munos, Alessandro Lazaric, and Mohammad Ghavamzadeh. Finite-sample analysis of Bellman residual minimization. In *Proceedings of the Second Asian Conference on Machine Learning (ACML)*, 2010. 16, 17, 21, 36, 75

Enno Mammen and Sara van de Geer. Locally adaptive regression splines. *The Annals of Statistics*, 25(1):387–414, 1997. 91

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)*. Omnipress, 2009. 74

Daniel McDonald. Generalization error bounds for state space models with an application to economic forecasting. Technical report, Department of Statistics, Carnegie Mellon University, July 2010. 133

Ron Meir. Nonparametric time series prediction through adaptive model selection. *Machine Learning*, 39(1):5–34, 2000. 49, 133

Francisco Melo, Sean P. Meyn, and Isabel Ribeiro. An analysis of reinforcement learning with function approximation. In Andrew McCallum and Sam Roweis, editors, *Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008)*, pages 664–671. Omnipress, 2008. 20

Ishai Menache, Shie Mannor, and Nahum Shimkin. Basis function adaptation in temporal difference reinforcement learning. *Annals of Operations Research*, 134(1):215–238, 2005. 18, 117

Shahar Mendelson. Lower bounds for the empirical risk minimization algorithm. *IEEE Transactions on Information Theory*, 54(8):3797 – 3803, August 2008. 160, 161

Sean Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, New York, NY, USA, 2009. 126, 136

Dharmendra S. Modha and Elias Masry. Memory-universal prediction of stationary random processes. *IEEE Transactions on Information Theory*, 44(1):117–133, 1998. 49, 133

Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for stationary ϕ-mixing and β-mixing processes. *Journal of Machine Learning Research*, 11:789–814, 2010. ISSN 1532-4435. 38

Abdelkader Mokkadem. Mixing properties of ARMA processes. *Stochastic Processes and their Applications*, 29(2):309 – 315, 1988. 37

David E. Moriarty, Alan C. Schultz, and John J. Grefenstette. Evolutionary algorithms for reinforcement learning. *Journal of Artificial Intelligence Research*, 11:241 – 276, 1999. 13

Rémi Munos. Error bounds for approximate policy iteration. In *Proceedings of the 20th Annual International Conference on Machine Learning (ICML)*, pages 560–567, 2003. 23, 24, 26, 28, 33, 35, 36, 74, 75

Rémi Munos. Performance bounds in L_p norm for approximate value iteration. *SIAM Journal on Control and Optimization*, pages 541–561, 2007. 23, 24, 31, 33, 34, 35, 36, 57, 66, 86, 133

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9:815–857, 2008. 15, 20, 21, 52, 53, 59, 60, 65, 66

Yavar Naddaf. Game-independent ai agents for playing atari 2600 console games. Master's thesis, Department of Computing Science, University of Alberta, April 2010. 2

Boaz Nadler, Nathan Srebro, and Xueyuan Zhou. Semi-supervised learning with the graph Laplacian: The limit of infinite unlabelled data. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems (NIPS - 22)*, pages 1330–1338, 2009. 113

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

Michael Nussbaum. Spline smoothing in regression models and asymptotic efficiency in l_2. *The Annals of Statistics*, 13(3):984–997, 1985. 160

Dirk Ormoneit and Saunak Sen. Kernel-based reinforcement learning. *Machine Learning*, 49:161–178, 2002. 19

Ronald Parr, Christopher Painter-Wakefield, Lihong Li, and Michael Littman. Analyzing feature generation for value-function approximation. In *ICML '07: Proceedings of the 24th international conference on Machine learning*, pages 737 – 744, New York, NY, USA, 2007. ACM. 18, 117

Ronald Parr, Lihong Li, Gavin Taylor, Christopher Painter-Wakefield, and Michael Littman. An analysis of linear models, linear value-function approximation, and feature selection for reinforcement learning. In *ICML '08: Proceedings of the 25th international conference on Machine learning*, pages 752–759. ACM, 2008. 18

Jan Peters, Vijayakumar Sethu, and Stefan Schaal. Reinforcement learning for humanoid robotics. In *Humanoids2003, Third IEEE-RAS International Conference on Humanoid Robots*, 2003. 13

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. 18, 19

Marek Petrik, Gavin Taylor, Ronald Parr, and Shlomo Zilberstein. Feature selection using regularization in approximate linear programs for markov decision processes. In Johannes Fürnkranz and Thorsten Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pages 871–878. Omnipress, 2010. 15

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

David Pollard. *Convergence of Stochastic Processes*. Springer Verlag, New York, 1984. 40

Leming Qu, Partha S. Routh, and Phil D. Anno. Wavelet reconstruction of nonuniformly sampled signals. *IEEE Signal Processing Letters*, 16(2):73 – 76, February 2009. 67

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over l_q-balls. Technical report, UC Berkely, Department of Statistics, October 2009. URL `arXiv:0910.2042v1`. 161

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated Gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259, 2010. 161

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. 56, 68, 118, 164

Martin Riedmiller. Neural fitted Q iteration – first experiences with a data efficient neural reinforcement learning method. In *16th European Conference on Machine Learning*, pages 317–328, 2005. 15, 19, 53

Jeffrey S. Rosenthal. *A Fist Look at Rigorous Probability Theory*. World Scientific Publishing, 2nd edition, 2006. 8

Stéphane Ross, Geoffrey Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In Geoffrey J. Gordon and David B. Dunson, editors, *Proceedings of the 14th International Conference on Artifical Intelligence and Statistics (AISTATS)*, pages 627–635, April 2011. 82

Paul-Marie Samson. Concentration of measure inequalities for Markov chains and ϕ-mixing processes. *The Annals of Probability*, 28(1):416–461, 2000. 82, 135, 136

Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002. 38, 164

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. 38, 55, 81, 92, 164

Dale Schuurmans and Relu Patrascu. Direct value-approximation for factored mdps. In *Advances in Neural Information Processing Systems (NIPS - 14)*, pages 1579–1586. MIT Press, 2001. 15

Paul J. Schweitzer and Abraham Seidmann. Generalized polynomial approximations in Markovian decision processes. *Journal of Mathematical Analysis and Applications*, 110: 568–582, 1985. 74

Clayton Scott and Robert Nowak. Minimax-optimal classification with dyadic decision trees. *IEEE Transactions on Information Theory*, 52:1335–1353, 2006. 162

Azad Shademan, Amir-massoud Farahmand, and Martin Jägersand. Robust Jacobian estimation for uncalibrated visual servoing. In *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, pages 5564–5569, May 2010. 1

John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, UK, 2004. 38, 52, 164

Bruno Siciliano and Oussama Khatib, editors. *Springer Handbook of Robotics*. Springer, 2008. 1, 148, 151

David Silver, Richard S. Sutton, and Martin Müller. Reinforcement learning of local shape in the game of go. In Manuela M. Veloso, editor, *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1053–1058, 2007. 2

Steve Smale and Ding-Xuan Zhou. Estimating the approximation error in learning theory. *Analysis and Applications*, 1(1):17–41, 2003. 45, 59, 84

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, 2008. 38, 44, 54, 58, 59, 83, 84, 113, 165

Ingo Steinwart, Don Hush, and Clint Scovel. Learning from dependent observations. *Journal of Multivariate Analysis*, 100(1):175–194, 2009. 37

Hongwei Sun and Qiang Wu. Regularized least square regression with dependent samples. *Advances in Computational Mathematics*, 32:175–189, 2010. 37

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning)*. The MIT Press, 1998. 7, 12, 17, 75, 115

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 *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pages 993–1000. ACM, 2009. 3, 17, 20

Csaba Szepesvári. The asymptotic convergence-rate of Q-learning. In *Advances in Neural Information Processing Systems (NIPS - 10)*, pages 1064–1070, 1997a. 20

Csaba Szepesvári. *Static and Dynamic Aspects of Optimal Sequential Decision Making*. PhD thesis, Bolyai Institute of Mathematics, University of Szeged, Szeged, Aradi vrt. tere 1, HUNGARY, 6720, September 1997b. 15

Csaba Szepesvári. *Algorithms for Reinforcement Learning*. Morgan Claypool Publishers, 2010. 4, 7, 17, 132

István Szita. Reinforcement learning in games. In Marco Wiering and Martijn van Otterlo, editors, *To appear in Reinforcement Learning: State of the Art*. Springer-Verlag, 2011. 2

Gavin Taylor and Ronald Parr. Kernelized value function approximation for reinforcement learning. In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1017–1024, New York, NY, USA, 2009. ACM. 16, 19, 91, 115

Gerald Tesauro. TD-gammon, a self-teaching backgammon program, achieves master-level play. *Neural Computation*, 6:215–219, 1994. 2

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996. 19, 67, 91, 161

Hans Triebel. *Theory of Function Spaces III*. Springer, 2006. 161

John N. Tsitsiklis and Benjamin Van Roy. An analysis of temporal difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42:674–690, 1997. 3, 17, 20

Alexander B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32 (1):135–166, 2004. 162

Sara A. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000. 39, 44, 58, 61, 82, 83, 92, 95, 97, 104, 105, 108, 109, 112, 138, 165

Sara A. van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009. 161

Aad W. van der Vaart, Sandrine Dudoit, and Mark J. van der Laan. Oracle inequalities for multi-fold cross validation. *Statistics and Decisions*, 24:351–372, 2006. 126

Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998. ISBN 0471030031. 19, 158

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010. 82

Mathukumalli Vidyasagar. *A Theory of Learning and Generalization: With Applications to Neural Networks*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, second edition, 2002. ISBN 1852333731. 37

Mathukumalli Vidyasagar and Rajeeva L. Karandikar. A learning theory approach to system identification and stochastic adaptive control. *Journal of Process Control*, 18(3–4):421 – 430, 2008. 37

Grace Wahba. *Spline Models for Observational Data*. SIAM [Society for Industrial and Applied Mathematics], 1990. 19, 38, 52, 54, 164

Larry Wasserman. *All of Nonparametric Statistics (Springer Texts in Statistics)*. Springer, 2007. 3, 18, 118, 158

Marten Wegkamp. Model selection in nonparametric regression. *The Annals of Statistics*, 31(1):252–273, 2003. 69, 116, 126

Shimon Whiteson and Peter Stone. Evolutionary function approximation for reinforcement learning. *Journal of Machine Learning Research*, 7(May):877–917, 2006. 115

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. 74

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. 16, 68, 90, 91

Xin Xu, Chunming Liu, and Dewen Hu. Continuous-action reinforcement learning with fast policy search and adaptive basis function selection. *Soft Computing*, March 2010. 68

Yong-Li Xu and Di-Rong Chen. Learning rates of regularized regression for exponentially strongly mixing sequence. *Journal of Statistical Planning and Inference*, 138(7):2180–2189, 2008. 37, 38

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. 68, 91

Yuhong Yang and Andrew R. Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999. 45, 65, 88, 158, 160

Yinyu Ye. The simplex and policy-iteration methods are strongly polynomial for the markov decision problem with a fixed discount rate. Technical report, Department of Management Science and Engineering, Stanford University, November 2010. 16

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 22(1):94–116, January 1994. 37, 38, 40, 82, 92

Huizhen Yu and Dimitri P. Bertsekas. Basis function adaptation methods for cost approximation in MDP. In *Proceedings of IEEE International Symposium on Adaptive Dynamic Programming and Reinforcement Learning*, pages 74 – 81, 2009. 18

Tong Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2(527 – 550), 2002. 67, 90

Tong Zhang. Some sharp performance bounds for least squares regression with l_1 regularization. *The Annals of Statistics*, 37(5A):2109–2144, 2009a. 67, 158

Tong Zhang. Adaptive forward-backward greedy algorithm for sparse learning with linear models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems (NIPS - 21)*, pages 1921–1928. 2009b. 67, 161

Ding-Xuan Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3): 739–767, 2002. 44, 58, 83, 165

Ding-Xuan Zhou. Capacity of reproducing kernel spaces in learning theory. *IEEE Transactions on Information Theory*, 49:1743–1752, 2003. 44, 58, 83, 165

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, December 2006. 67, 161

Hui Zou and Trevor Hastie. Regularization and variable selection via elastic net. *Journal of the Royal Statistical Society, Series B*, 67(2):301 – 320, 2005. 67

Appendix A

Supervised Learning

The supervised learning literature provides two types of insights about the RL/Planning problems:

- A lower bound for a regression problem leads to a lower bound for a related policy evaluation problem.
- The powerful supervised learning algorithms may inspire us to design new algorithms for the RL/Planning problems with large state spaces.

One can easily see that the regression problem is indeed a special case of the policy evaluation problem when the discount factor γ is set to zero. Consequently a lower bound for the regression problem is also a lower bound for the subproblem of policy evaluation in the RL/Planning context. In Section A.1, we survey some lower bounds for the regression problem. We quote a result that shows that if the problem does not have any well-behaving regularity, the learning process can be arbitrary slow. On the other hand if the problem has some kind of regularities, learning becomes feasible and we may have a reasonable convergence rate. These regularities are usually measured according to some notion of complexity of the problem. Examples of complexity measures are the Vapnik-Chervonenkis (VC) dimension [Vapnik, 1998], various notions of smoothness [Györfi et al., 2002], metric entropy [Yang and Barron, 1999], the degree of sparsity [Lafferty and Wasserman, 2006; Zhang, 2009a], and the global and local Rademacher complexities [Bartlett and Mendelson, 2002; Koltchinskii, 2006] of the function (hypothesis) space to which the target function belongs. Section A.2 is devoted to various types of regularities that are well-studied in the supervised learning literature.

Another way that the supervised learning literature can help is as the source of inspiration to design new RL/Planning algorithms. The existence of many flexible and adaptive supervised learning algorithms encourages us to adopt them for RL/Planning problems. The focus of this work has been the regularization-based algorithms as described in Chapters 5, 6, and 7. Evidently, the regularization-based algorithms are not the only powerful class of algorithms in the supervised learning literature, and one may expect to design new algorithms based on other powerful techniques too. The literature on supervised learning is abundant and we do not even attempt to review them here. Instead, we refer the reader to standard textbooks such as Hastie et al. [2001]; Bishop [2006] for the comprehensive coverage of the supervised learning algorithms and Devroye et al. [1996]; Györfi et al. [2002]; Wasserman [2007] for theoretical analyses of them.

A.1 Lower Bounds for the Regression Problem

The lower bounds or *slow rates* provide insight about the intrinsic difficulty of learning problems. They show how many samples are required in the worst case to estimate the

regression/classifier/density/value function up to a specific accuracy. These results are interesting because they demonstrate the intrinsic difficulty of learning problems – as opposed to the performance of a particular algorithm. Briefly speaking, the available lower bounds indicate that learning can be hopelessly difficult, unless there are some intrinsic regularities in the problem. If there are, an estimator exists that has a rate of convergence, which itself depends on some well-specified notion of the problem's complexity. In this section, we present some particular examples of lower bound results for the regression problem and in Section A.2 we discuss different types of regularities from a higher-level viewpoint.

First let us briefly formalize the regression problem, which will be the focus of the rest of this section. Consider a pair of random variables (X, Y) where $X \in \mathcal{X}$ and $Y \in \mathbb{R}$ with the joint probability distribution μ_{XY} (or simply μ). Also assume that we are given a dataset $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ with $(X_i, Y_i) \sim \mu_{XY}$. The *regression function* is a real-valued function with the domain $\mathcal{X}$ and is defined as $m(x) = \mathbb{E}[Y|X = x]$. It can be shown that this function is the minimizer of the L_2-risk, i.e., $m \leftarrow \operatorname{argmin}_f \|f(X) - Y\|_\mu^2$.

When we do not know the joint distribution μ_{XY}, which is usually the case in practice, we cannot analytically determine the regression function m. Instead, we use samples $\mathcal{D}_n$ to estimate the function $\hat{m}_n(\cdot; \mathcal{D}_n) : \mathcal{X} \to \mathbb{R}$. The goal is to have an estimate $\hat{m}_n(\cdot; \mathcal{D}_n)$ that has a small *excess error* $\|\hat{m}_n(\cdot; \mathcal{D}_n) - m\|$. The following negative result shows that the regression problem can be arbitrary difficult.

Negative Result

Theorem A.1 (Györfi et al. [2002] – Theorem 3.1)**.** *Let $\{a_n\}$ be a sequence of positive numbers converging to zero. For every fixed sequence of regression estimates $\{\hat{m}_n(\cdot; \mathcal{D}_n)\}$, there exists a distribution μ_{XY}, such that X is uniformly distributed on $[0, 1]$, $Y = m(X) = \pm 1$, and*

$$\limsup_{n\to\infty} \frac{\mathbb{E}\left[\|\hat{m}_n - m\|^2\right]}{a_n} \geq 1.$$

This theorem states that for a subset of all regression problems, where X is distributed uniformly and Y is noiseless samples that can be either $+1$ or -1, the convergence rate can be arbitrary slow. This result indicates that we cannot hope to have any estimator that has a convergence rate for all problems – even if the distribution of X is known and Ys are noiseless.

Nevertheless, if we restrict the range of problems μ_{XY} to a small subset of joint distributions with certain amount of structure/regularities, we can get a convergence rate. In the rest of this section, we provide several examples of such results. The difference between these examples is in the way the regularity is defined. First, we cover the *smoothness* regularities, then we provide a result when the regularity is defined according to the *metric packing entropy* of the function space (Definition B.6 in Appendix B.2), and finally we cite a result regarding the influence of the *geometry* of the function space on the convergence rate.

Let us define the class of (p, C)-smooth functions [Györfi et al., 2002, Chapter 3].

Smoothness Lower Bound

Definition A.1 ((p, C)-smoothness)**.** *Let $p = k + \beta$ for some $k \in \mathbb{N}_0$ and $0 < \beta \leq 1$, and let $C > 0$. A function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ is called (p, C)-smooth if for every $\alpha = (\alpha_1, \ldots, \alpha_d)$ ($\alpha_i \in \mathbb{N}_0$, $\sum_{i=1}^d \alpha_i = k$) the partial derivative $\frac{\partial^k f}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}$ exists and for all $x, y \in \mathbb{R}^d$ satisfies*

$$\left| \frac{\partial^k f}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}(x) - \frac{\partial^k f}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}(z) \right| \leq C \|x - z\|^\beta .$$

Define $\mathcal{F}^{(p,C)}$ to be the set of all (p, C)-smooth functions $f : \mathbb{R}^d \longrightarrow \mathbb{R}$.

Let us define the following class of regression problems.

Definition A.2 (Class of $\mathcal{D}^{(p,C)}$ Problems)**.** *Let $\mathcal{D}^{(p,C)}$ be the class of regression problems such that*

- X *is uniformly distributed on* $[0,1]^d$,
- $Y = m(X) + \eta$, *where* X *and* η *are independent and* η *is a standard normal random variable,*
- $m \in \mathcal{F}^{(p,C)}$.

We have the following lower bound.

Theorem A.2 (Minimax and Individual Lower Bounds for $\mathcal{D}^{(p,C)}$ –Antos et al. [2000]). *For the class* $\mathcal{D}^{(p,C)}$, *there exists some constant* B *independent of* C *such that*

$$\liminf_{n\to\infty} \inf_{\hat{m}_n} \sup_{(X,Y)\in\mathcal{D}^{(p,C)}} \mathbb{E}\left[\|\hat{m}_n - m\|^2\right] \geq BC^{\frac{2d}{2p+d}} n^{-\frac{2p}{2p+d}},$$

This is called the minimax lower bound *of convergence. Moreover, consider* $\{b_n\}$ *as an arbitrary positive sequence tending to zero. We have*

$$\inf_{\hat{m}_n} \sup_{(X,Y)\in\mathcal{D}^{(p,C)}} \limsup_{n\to\infty} \mathbb{E}\left[\|\hat{m}_n - m\|^2\right] > b_n n^{-\frac{2p}{2p+d}}.$$

which is called the individual lower bound *of convergence.*

As another related example, Nussbaum [1985] considers the regression problem with fixed design. He provides an optimal minimax rate with sharp constants when the regression function belongs to $\{f : f \in \mathbb{W}^k(\mathbb{R}^d), \|D^m f\|^2 \leq J_0^2\}$.

Metric Entropy Lower Bound

One may also provide a lower bound for learning when the complexity of the function space is described according to its metric entropy. Consider the regression model $Y_i = m(X_i) + \eta_i$ $(i = 1, \ldots, n)$ with η_is are i.i.d. with the standard normal distribution and X_is are also i.i.d. with the distribution μ_X. Let $\mathcal{M}(\varepsilon, \mathcal{F}, \|\cdot\|_{2,\mu_X})$ be the ε-packing number of $\mathcal{F}$ w.r.t. $\|\cdot\|_{2,\mu_X}$.

Theorem A.3 (Minimax Lower Bound for the Class of Controlled Metric Packing Entropy – Theorem 6 of Yang and Barron [1999]). *For the class of bounded functions* $\mathcal{F}$, *assume that for some* $0 < \rho < 1$,

$$\liminf_{\varepsilon\to 0} \frac{\mathcal{M}\left(\rho\,\varepsilon, \mathcal{F}, \|\cdot\|_{2,\mu_X}\right)}{\mathcal{M}\left(\varepsilon, \mathcal{F}, \|\cdot\|_{2,\mu_X}\right)} > 1.$$

Choose ε_n *such that* $\log \mathcal{M}(\varepsilon_n, \mathcal{F}, \|\cdot\|_{2,\mu_X}) = n\varepsilon_n^2$. *Then*

$$\inf_{\hat{m}_n} \sup_{m\in\mathcal{F}} \mathbb{E}\left[\|\hat{m}_n - m\|^2\right] = \Theta(\varepsilon_n^2).$$

This theorem is relevant to our results in which the capacity is described according to the metric entropy condition (cf. Assumptions A2, A7, and A16). According to this theorem, a capacity condition in the form of $\mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|_{2,\mu_X}) = \varepsilon^{-2\alpha}$ (for a fixed $0 < \alpha < 1$ and any $\varepsilon > 0$), implies a minimax optimal convergence rate of $\Theta(n^{-\frac{1}{1+\alpha}})$ for the squared error, which is the same rate as our upper bounds in Chapters 4, 5, 6.

Geometric Lower Bound

Some results are geometric flavored. They indicate that if the function space $\mathcal{F}$ from which the estimator $\hat{m}_n(\cdot; \mathcal{D}_n)$ is picked has a "bad" geometry, the learning cannot be very fast. The notions of badness can be defined in different ways. For instance, Lee et al. [1998, 2008] show that if the closure of a finite dimensional function space $\mathcal{F}$ is *not convex*, the L_2-loss of any learning algorithm cannot converge to the best possible function in $\mathcal{F}$ faster than $O(\frac{1}{\sqrt{n}})$. The following theorem due to Mendelson [2008] is a modification of the result of Lee et al. [1998]

Theorem A.4 (Mendelson [2008]). *Let $\mathcal{F} \subset L_2(\mu_{\mathcal{X}})$ be a compact class of functions bounded by 1. Assume that there is a random variable Y bounded by B for which $\inf_{f\in\mathcal{F}} \mathbb{E}\left[|Y-f|^2\right]$ has more than a unique minimizer in $\mathcal{F}$. Consider a learning algorithm that for every integer n and any sample $\mathcal{D}_n$ assigns a function $\hat{m}_n(\cdot;\mathcal{D}_n) \in \mathcal{F}$. Then there are constants c and n_0 depending only on $\mathcal{F}$, B, and $\mu_{\mathcal{X}}$ such that for any $n \geq n_0$,*

$$\sup_{Y\in\{Y:\|Y\|_\infty\leq B\}} \mathbb{E}\left[|Y-\hat{m}_n(X;\mathcal{D}_n)|^2\right] - \inf_{f\in\mathcal{F}} \mathbb{E}\left[|Y-f(X)|^2\right] \geq \frac{c}{\sqrt{n}}.$$

The difference between these results and Theorems A.2 and A.3 is that the source of difficulty here is not because of the richness of the function space but is the consequence of its bad geometry.

A.2 On Regularities

The results of Section A.1 show that solving a regression problem might not be possible unless there is some underlying regularities in the problem. In that case, it is desirable to have an *adaptive* algorithm that automatically detects the present regularity and exploits it. In this section we provide a high-level overview of the following common types of regularities studied in the statistics/supervised learning literature:

- Smoothness
- Sparsity
- Low-Dimensionality of the Input Manifold
- Low-Noise Margin Condition

Smoothness

Smoothness of the target function is one of the most common ways to describe the regularity of a problem. There are various notions of smoothness such as Hölderian smoothness that requires the derivatives of the function to be Hölder continuous (Definition B.2 in Appendix B.2) and the smoothness according to the Sobolev norm that requires the weak-derivatives of the function to be L_p-integrable (Definition B.3 in Appendix B.2). This latter notion of smoothness allows the function to have occasional discontinuities. For a comprehensive treatment of various notions of smoothness, refer to Triebel [2006, Chapter 1: How to Measure Smoothness], and for some typical results on the statistical behavior of estimators under these conditions, refer to Györfi et al. [2002].

Sparsity

Sparsity is another type of regularity that has recently attracted considerable attention [Tibshirani, 1996; Donoho and Johnstone, 1995; Zou, 2006; Zhang, 2009b; Lafferty and Wasserman, 2006]. Consider a p-dimensional function space $\mathcal{F}$ with $\{\Phi_i\}_{i=1}^p$ as its basis functions, i.e., any function $f \in \mathcal{F}$ has an expansion $f(\cdot) = \sum_{i=1}^p w_i \Phi_i(\cdot)$ for $w \in \mathbb{R}^p$. A function f is said to be s-sparse when the number of non-zero elements of w is s, i.e., $s = |\{w_i \neq 0 : i = 1, \ldots, p\}|$. Sparsity of the target function allows us to design more efficient estimation procedures. If the true s-sparse function has the parameter vector w^*, one can show that under certain conditions on the design matrix, the parameter estimation error $\|\hat{w} - w^*\|^2$ of LASSO [Tibshirani, 1996] would be $O(\frac{s\log(p)}{n})$. If p is comparable to n (or even $p \gg n$) and the target function is s-sparse with $s \ll n$, the improvement over $O(\frac{p}{n})$ behavior of a procedure that does not exploit the sparsity is notable. For the review on the conditions that allow a procedure such as LASSO to achieve such a rate, see e.g., van de Geer and Bühlmann [2009]; Raskutti et al. [2010]. Raskutti et al. [2009] provides minimax convergence rates for estimation when w^* belongs to an l_q-ball with $q \in [0, 1]$.

Low-dimensional Data Manifold

Low-dimensionality of data manifold is a geometrical regularity describing the situation that input data belongs to a D-dimensional space $\mathcal{X}$ but they are confined (or close) to a d-dimensional manifold $\mathcal{M} \subset \mathcal{X}$. We call an algorithm *manifold-adaptive* if it exploits this

property and performs as if the dimension of the input space is d. This leads to a huge statistical performance gain whenever $d \ll D$.

Recently, there have been a few theoretical results that show the possibility of having manifold-adaptive algorithms. Farahmand et al. [2007b] shows that the sample complexity of estimating the dimension of manifold $\mathcal{M}$ depends mainly on the intrinsic dimension d of $\mathcal{M}$ and not the dimension D of the embedding space $\mathcal{X}$. Farahmand et al. [2007c] present a result that shows that a simple K-nearest neighborhood regression algorithm is also manifold-adaptive (Note that the conventional K-nearest neighborhood-based algorithms do not exploit other regularities of the problem such as its smoothness). See the work of Farahmand et al. [2011] for the detailed discussion of this result. Among other works that prove manifold-adaptivity of a procedure, we can refer to Scott and Nowak [2006] that introduces dyadic decision trees for classification. Another approach with favorable manifold-adaptive properties is Random Projection Tree (Dasgupta and Freund [2008]) which is a variant of k-d trees. It uses random splitting directions instead of splitting along a coordinate direction and uses the randomly-perturbed median as the point of splitting. Nevertheless, to best of our knowledge, we are far from a general statistical theory of manifold-adaptive algorithms.

Low-Noise Margin

Low-noise margin condition is another type of regularity that appears in classification problems. This condition is regarding the behavior of *a posteriori probability* function $\eta(x) = \mathbb{P}\{Y = 1 | X = x\}$ around the critical decision point $\frac{1}{2}$ (for 0/1-classification problem). It turns out that if $\eta(x)$ is far away from $\frac{1}{2}$ for most $x \in \mathcal{X}$, one can show improved convergence rates for the classification problem, see e.g., [Tsybakov, 2004; Audibert and Tsybakov, 2007] and Section 5.2 of Boucheron et al. [2005]. The quantitative behavior of $\eta(x)$ around $\frac{1}{2}$ can be described by different conditions such as the *Massart* or *Tsybachov* noise conditions.

Our short discussion here about different types of regularities should not imply that these are the only possible regularities one may exploit in any problem. There are several other types of regularities that explicitly or implicitly have been studied in the machine learning and statistics literature (e.g., ANOVA decomposability). Moreover, one can be sure that there will be several undiscovered regularities in real-world learning problems that might be useful to consider when designing new algorithms. One may also speculate that real-world RL/Planning problems have regularities that do not come up in the supervised learning problems. Discovering and studying them should be the subject of future research.

Appendix B

Mathematical Background

In this appendix, we briefly review some mathematical definitions and results that are used in the thesis.

B.1 Function Spaces

Definition B.1. *For $\mathcal{X}$ an open subset of $\mathbb{R}^d$, a function $f : \mathcal{X} \to \mathbb{R}$ is Hölder continuous if for some nonnegative finite real numbers C and α,*

$$|f(x) - f(y)| \leq C|x - y|^{\alpha}. \qquad (x, y \in \mathcal{X})$$

The value of α is called the exponent of the Hölder condition.

For example, when $\alpha = 1$, the Hölder continuity is the same as Lipschitz continuity; and $\alpha = 0$ implies that the function is bounded.

Definition B.2. *Let k be a nonnegative integer number and $0 < \alpha \leq 1$. The Hölder space $C^{\alpha,k}(\mathcal{X})$ is the space of all functions with domain $\mathcal{X}$ that have derivatives up to order k and their k^{th} partial derivatives are Hölder continuous with exponent α.*

Definition B.3 (Sobolev Space $\mathbb{W}^{k,p}(\mathcal{X})$ – Devore [1998])**.** *Let k be a nonnegative integer number and $1 \leq p \leq \infty$. The Sobolev space $\mathbb{W}^{k,p}(\mathcal{X})$ for open and connected subset $\mathcal{X}$ of $\mathbb{R}^d$ is the space of all measurable functions whose distributional derivative of order k is in $L_p(\mathcal{X})$, i.e.,*

$$\left\| \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} \right\|_{L_p(\mathcal{X})} < \infty,$$

for every multi-index $|\alpha| \leq k$. The semi-norm for $\mathbb{W}^{k,p}(\mathcal{X})$ is defined as

$$|f|_{\mathbb{W}^{k,p}(\mathcal{X})} \triangleq \sum_{|\alpha|=k} \left\| \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}} \right\|_{L_p(\mathcal{X})},$$

and their norm by $\|f\|_{\mathbb{W}^{k,p}(\mathcal{X})} \triangleq |f|_{\mathbb{W}^{k,p}(\mathcal{X})} + \|f\|_{L_p(\mathcal{X})}$. Denote $\mathbb{W}^{k,2}(\mathbb{R}^d)$ by $\mathbb{W}^k(\mathbb{R}^d)$.

Sobolev spaces generalize Hölder spaces by allowing functions that are only almost everywhere differentiable. Another relevant class of function spaces is the class of Besov spaces $\mathcal{B}^s_{p,q}(\mathcal{X})$ for $0 < p, q \leq \infty$ and $s > 0$. Besov spaces generalize Sobolev spaces by letting $0 < p < 1$ and having fractional smoothness order s. For instance, $\mathcal{B}^s_{2,2}(\mathbb{R}^d)$ is the same as $\mathbb{W}^{s,2}(\mathbb{R}^d)$. We do not define Besov spaces here, and only mention that Besov spaces can be defined with the help of *modulus of smoothness*. See Devore [1998] for more information.

One may extend this definition to domain $\mathcal{X} \times \mathcal{A}$ with a finite $\mathcal{A}$ as well. First, define the distributional derivative of order k of $Q \in \mathcal{F}^{|\mathcal{A}|}$ by

$$\frac{\partial^{|\alpha|} Q}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} = \begin{bmatrix} \frac{\partial^{|\alpha|} Q(x,a_1)}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} \\ \cdots \\ \frac{\partial^{|\alpha|} Q(x,a_{|\mathcal{A}|})}{\partial x_1^{\alpha_d} \cdots \partial x_d^{\alpha_d}} \end{bmatrix}.$$

Also define the $L_2(\mathcal{X} \times \mathcal{A})$-norm of Q as $\|Q\|_{L_2(\mathcal{X}\times\mathcal{A})} = \sum_{a\in\mathcal{A}} \|Q(\cdot, a)\|_{L_2(\mathcal{X})}$. This leads to $\|Q\|^2_{\mathbb{W}^k(\mathbb{R}^d\times\mathcal{A})} = \sum_{a\in\mathcal{A}} \|Q(\cdot, a)\|^2_{\mathbb{W}^k(\mathbb{R}^d)}$.

B.1.1 Reproducing Kernel Hilbert Spaces

The following definition of an RKHS is borrowed from Aronszajn [1950].

Definition B.4 (Aronszajn [1950]). *Let $\mathcal{H}$ be a Hilbert space defined in $\mathcal{X}$ with the inner product $\langle \cdot, \cdot \rangle$. The function $\kappa(x, y)$ of x and y in $\mathcal{X}$ is called a **reproducing kernel** of $\mathcal{F}$ if*

- *For every $y \in \mathcal{X}$, $\kappa(\cdot, y) \in \mathcal{H}$.*
- *For every $y \in \mathcal{X}$, and for every $f \in \mathcal{H}$, we have $f(y) = \langle f(x), \kappa(x, y) \rangle$.*

We quote the following slight generalization of Schölkopf et al. [2001, Theorem 1]).

Theorem B.1 (Generalized Representer Theorem). *Let $\Omega : [0, \infty) \to \mathbb{R}$ be a strictly monotonically increasing function, $\mathcal{X}$ be a set, $\mathcal{H}$ be an RKHS with kernel $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and $c_n : \mathcal{X}^n \to \mathbb{R}$ be an arbitrary loss function. Then any $f \in \mathcal{H}$ minimizing*

$$c_n\big(f(x_1), \ldots, f(x_n)\big) + \Omega\big(\|f\|_{\mathcal{H}}\big)$$

admits a representation of the form $f(x) = \sum_{t=1}^n \alpha_t k(x_t, \cdot)$.

For more information on RKHSs and their properties, see e.g., Aronszajn [1950], Wahba [1990], Schölkopf and Smola [2002], Cucker and Smale [2002, Chapter III], Shawe-Taylor and Cristianini [2004], and Rasmussen and Williams [2006, Section 6.1].

B.2 Covering Number and Metric Entropy

The following definitions are from Györfi et al. [2002, Chapter 9].

Definition B.5 (Covering Number – 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 $\mathcal{X}$.*

1. *Every finite collection of $N_\varepsilon = \{f_1, \ldots, f_{N_\varepsilon}\}$ 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. $\|\cdot\|_{p,\nu_{\mathcal{X}}}$.*
2. *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 $x_{1:n} = (x_1, \ldots, x_n) \in \mathcal{X}^n$, 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 $\|\cdot\|_{p,n}$ and is denoted by $\mathcal{N}_p(\varepsilon, \mathcal{F}, x_{1:n})$ (or occasionally $\mathcal{N}_p(\varepsilon, \mathcal{F}, x_1^n)$). If $X_{1:n} = (X_1, \ldots, X_n)$ is a sequence of random variables, the covering number $\mathcal{N}_p(\varepsilon, \mathcal{F}, X_{1:n})$ is a random variable too.

A related concept is the *packing number* of a class of functions $\mathcal{F}$.

Definition B.6 (Packing Number – Definition 9.4 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 $\mathcal{X}$.*

1. *A finite set $M_\varepsilon = \{f_1, \ldots, f_{M_\varepsilon}\}$ is said to be an ε-packing of $\mathcal{F}$ w.r.t. $\|\cdot\|_{p,\nu_{\mathcal{X}}}$ if for any $f_i, f_j \in M_\varepsilon$ $(f_i \neq f_j)$, we have $\|f_i - f_j\|_{p,\nu_{\mathcal{X}}} > \varepsilon$.*

2. *Let $\mathcal{M}(\varepsilon, \mathcal{F}, \|\cdot\|_{p,\nu_{\mathcal{X}}})$ be the size of the largest ε-packing of $\mathcal{F}$ w.r.t. $\|\cdot\|_{p,\nu_{\mathcal{X}}}$. If for every $M \in \mathbb{N}$, there exists an ε-packing of $\mathcal{F}$ w.r.t. $\|\cdot\|_{p,\nu_{\mathcal{X}}}$ with size M, then take $\mathcal{M}(\varepsilon, \mathcal{F}, \|\cdot\|_{p,\nu_{\mathcal{X}}}) = \infty$. Then $\mathcal{M}(\varepsilon, \mathcal{F}, \|\cdot\|_{p,\nu_{\mathcal{X}}})$ is called an ε-**packing number** of $\mathcal{F}$ w.r.t. $\|\cdot\|_{p,\nu_{\mathcal{X}}}$.*

In this work, we often refer to the covering number and the metric entropy of a class of functions. Occasionally, however, we may refer to the packing number of a class of functions. In these cases, we use *metric packing entropy number* to refer to the logarithm of the packing number.

The following is an example upper bound on the metric entropy of certain classes of RKHSs.

Proposition B.2 (Theorem 4 – Zhou [2003]). *let $K : [0,1]^d \times [0,1]^d \to \mathbb{R}$ be a Mercer kernel. If $s > 0$ and K lies in $Lip^*(s, \mathcal{C}([0,1]^d \times [0,1]^d))$, then*

$$\log \mathcal{N}_2(u, \overline{I_K(\mathcal{B}_R)}) \leq C \left(\frac{R}{u}\right)^{\frac{2d}{s}},$$

where $I_K(\mathcal{B}_R)$ is the inclusion of $\mathcal{B}_R$ in $\mathcal{C}(X)$, the space of continuous functions, and $\overline{A}$ is the closure of A.

Refer to Györfi et al. [2002]; van de Geer [2000]; Zhou [2002, 2003] for some other examples.

B.3 Peeling Device

The following definition of the peeling device is from Section 5.3 of van de Geer [2000].

Consider the function space $\mathcal{F}$ and let $X_n(f)$ be an appropriately-defined stochastic process indexed by $\mathcal{F}$. Consider the function $\tau : \mathcal{F} \to [\rho, \infty)$ $(\rho > 0)$. The goal is to have a probability upper bound on the weighted process $|X_n(f)|/\tau(f)$.

Let $(\sigma_l)_{l \geq 0}$ be a strictly increasing sequence with $\sigma_0 = 0$ and $\lim_{l \to \infty} \sigma_l = \infty$. The function space $\mathcal{F}$ can be "peeled" off into the following "smaller" function spaces:

$$\mathcal{F} = \bigcup_{l \geq 1} \mathcal{F}_{\sigma_l}$$

with $\mathcal{F}_{\sigma_l} \triangleq \{ f \in \mathcal{F} : \sigma_{l-1} \leq \tau(f) < \sigma_l \}$ $(l = 1, 2, \ldots)$. For any positive a, we have

$$\mathbb{P}\left\{ \sup_{f \in \mathcal{F}} \frac{|X_n(f)|}{\tau(f)} > a \right\} \leq \sum_{l \geq 1} \mathbb{P}\left\{ \sup_{f \in \mathcal{F}_{\sigma_l}} \frac{|X_n(f)|}{\tau(f)} > a \right\} \leq \sum_{l \geq 1} \mathbb{P}\left\{ \sup_{f \in \mathcal{F}, \tau(f) < \sigma_l} |X_n(f)| > a\, \sigma_{l-1} \right\}.$$

This procedure is called the *peeling device*; and each $l = 1, 2, \ldots$ denotes a layer of peeling.

B.4 Carathéodory Sets

The definition and properties of Carathéodory sets are borrowed from Section 7.3 of Steinwart and Christmann [2008].

Let (T, d) be a metric space and $(\mathcal{X}, \sigma_{\mathcal{X}})$ be a measurable space. A family of measurable maps $(f_t)_{t \in T}$ is called a Carathéodory family if $t \mapsto f_t(x)$ is continuous for all $x \in \mathcal{X}$. Moreover, if T is separable or complete, we say that $(f_t)_{t \in T}$ is separable or complete,

respectively. A measurable set $\mathcal{F}$ on $\mathcal{X}$ is (separable or complete) Carathéodory set if there exists a (separable or complete) metric space (T, d) and a Carathéodory family $(f_t)_{t\in T}$ such that $\mathcal{F} = \{ f_t : t \in T \}$. A Carathéodory set $\mathcal{F}$ satisfies

$$\sup_{f\in\mathcal{F}} f(x) = \sup_{t\in T} f_t(x) = \sup_{t\in S} f_t(x)$$

for all dense $S \subset T$. For a separable Carathéodory set $\mathcal{F}$, the map $x \mapsto \sup_{t\in T} f_t(x)$ is measurable. Also for a separable and complete Carathéodory set $\mathcal{F}$, the map $(x, t) \mapsto f_t(x)$ is measurable.

B.5 Fixed-Point Theorem

Theorem B.3 (Banach Fixed-Point Theorem – Hutson et al. [2005])**.** *Let $(\mathcal{X}, d)$ be a non-empty complete metric space. Let $\mathcal{L} : \mathcal{X} \to \mathcal{X}$ be a contraction mapping on $\mathcal{X}$. Then the map $\mathcal{L}$ admits a unique fixed point $f^* = \mathcal{L}f^*$ with $f^* \in \mathcal{X}$. The fixed point can be found by the iterative application of $\mathcal{L}$ on arbitrary $f_0 \in \mathcal{X}$, i.e., $f^* = \lim_{k\to\infty} \mathcal{L}^k f_0$.*