CSIAM Trans. Appl. Math.
doi: 10.4208/csiam-am.SO-2021-0026

Vol. **3**, No. 2, pp. 191-220
June 2022

An L^2 Analysis of Reinforcement Learning in High Dimensions with Kernel and Neural Network Approximation

Jihao Long[1], Jiequn Han[2,3,*] and Weinan E[4,2,1]

[1] *Program of Applied and Computational Mathematics, Princeton University, Princeton, 08544, USA.*
[2] *Department of Mathematics, Princeton University, Princeton, 08544, USA.*
[3] *Center for Computational Mathematics, Flatiron Institute, New York, 10010, USA.*
[4] *School of Mathematical Sciences, Peking University, Beijing, 100871, China.*

Received 14 April 2021; Accepted 6 November 2021

Abstract. Reinforcement learning (RL) algorithms based on high-dimensional function approximation have achieved tremendous empirical success in large-scale problems with an enormous number of states. However, most analysis of such algorithms gives rise to error bounds that involve either the number of states or the number of features. This paper considers the situation where the function approximation is made either using the kernel method or the two-layer neural network model, in the context of a fitted Q-iteration algorithm with explicit regularization. We establish an $\tilde{O}(H^3|\mathcal{A}|^{\frac{1}{4}}n^{-\frac{1}{4}})$ bound for the optimal policy with Hn samples, where H is the length of each episode and $|\mathcal{A}|$ is the size of action space. Our analysis hinges on analyzing the L^2 error of the approximated Q-function using n data points. Even though this result still requires a finite-sized action space, the error bound is independent of the dimensionality of the state space.

AMS subject classifications: 68Q25, 62R07, 68T07, 93C55, 93C57

Key words: Reinforcement learning, function approximation, neural networks, reproducing kernel Hilbert space.

1 Introduction

Modern reinforcement learning (RL) algorithms often deal with problems involving an enormous amount of states, often in high dimensions, where function approximation must be introduced for the value or policy functions. Despite their practical success [20,

*Corresponding author. *Email addresses:* jihaol@princeton.edu (J. Long), jiequnhan@gmail.com (J. Han), weinan@math.princeton.edu (W. E)

40,49], most existing theoretical analysis of RL is only applicable to the tabular setting (see e.g. [3, 4, 17, 32, 33, 43]), in which both the state and action spaces are discrete and finite, and the value function is represented by a table without function approximation. Relatively simple function approximation methods, such as the linear model [34, 57] or generalized linear model [55], have been studied in the context of RL with various statistical estimates. The kernel method has also been studied in [18, 28, 60, 61], but results therein either suffer from the curse of dimensionality or require stringent assumptions about the kernel (in the form of fast decay of the kernel's eigenvalues or the bounds on the covering number). This paper considers general kernel method and two-layer neural network models and establishes dimension-independent results for these two classes of function approximation.

In the context of supervised learning, dimension-independent error rates have been established for a number of important machine learning models [21], including the kernel methods and two-layer neural network models. Of particular importance is the choice of the function space associated with the specific machine learning model. For the kernel method and two-layer neural network models, the corresponding function spaces are the reproducing kernel Hilbert space (RKHS) [1] and Barron space [22], respectively. Extending such results to the setting of reinforcement learning is a challenging task due to the coupling between the value/policy functions at different time steps.

In this work, we consider the fitted Q-iteration algorithm [13, 26, 27, 46, 51] for situations where the state space is embedded in a high-dimensional Euclidean space and the action space is finite. The function approximation to the Q-function is made either using the kernel method or the two-layer neural networks, with explicit regularization. We assume there is a *simulator* that can generate samples for the next state and reward, given the current state and the action. This allows us to focus on analyzing errors from the function approximation. Under the assumptions that the function approximation is compatible with the reward function and transition model, and all admissible distributions are uniformly bounded from a reference distribution (see Assumption 3.2), we establish an $\tilde{O}(H^3|\mathcal{A}|^{\frac{1}{4}}n^{-\frac{1}{4}})$ bound for the optimal policy with Hn samples, where H is the length of each episode and $|\mathcal{A}|$ is the size of action space. This result is independent of the dimensionality of the state space, and the convergence rate for n is close to the statistical lower bound for many function spaces, including several popular cases of RKHS and the Barron space (see Section 5 for a detailed discussion).

The key component in the analysis is to estimate the one-step error and control the error propagation. An important issue is the choice of the norm. L^∞ estimates have been popular in reinforcement learning for analyzing the tabular setting [4, 17, 32, 33], linear models, [11, 34, 57] and kernel methods [60, 61] (see discussions below). However, in the case we are considering, L^∞ estimates suffer from the curse of dimensionality with respect to the sample complexity, i.e. to ensure that the error is smaller than ϵ, we need at least $O(\epsilon^{-d})$ samples in d-dimensional state space (see Section 5 for a detailed discussion). This fact also explains why we only consider finite action space. Once we consider the high-dimensional action space, it is inevitable to find a maximum of a high-dimensional