On the role of overparameterization in Temporal Difference learning with linear function approximation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Much of the recent successes of deep learning can be attributed to scaling up the size of the networks to the point where they often are vastly overparameterized. Thus, understanding the role of overparameterization is of increasing importance. While predictive theories have been developed for supervised learning, little is known about the Reinforcement Learning case. In this work, we take a theoretical approach and study the role of overparameterization for Temporal Difference (TD) learning in the linear setting. We leverage tools from random matrix theory and random graph theory to obtain a characterization of the spectrum of the TD operator. We use this result to study the stability and optimization dynamics of TD learning as a function of the number of parameters.

1 Introduction

Occam's razor principle states that among many plausible explanations, the simplest one is the most likely to be true. In statistics and machine learning, this is often interpreted as models with a restricted number of parameters should be privileged as they generalize better [1]. This view is supported by the classical *bias-variance trade-off* which implies that models that are too flexible are bound to overfit [2]. Yet, deep neural networks with more parameters than training data points achieved remarkable performance in many domains [3, 4, 5], challenging this conventional wisdom. A significant understanding of this apparent tension has been reached in recent years. A large body of work using tools from *random matrix theory* has shown that, in models with random features, overparameterization does not lead to overfitting but instead better generalization [6]. This phenomenon is known as *double descent* and unifies the traditional and modern understanding of bias and variance. Importantly, [7] has shown that the double descent phenomenon is in large part explained by purely optimization rather than label noise. Following their insight, we mainly take an optimization approach in the rest of our paper.

While we now have a theory of overparameterization in supervised learning, we are not aware of any theoretical work on overparameterization in the reinforcement learning (RL) setting. Indeed, RL poses several new additional challenges. First, the agent has to interact with an *environment* whose structure may be complex. Second, each state may not sampled with equal probability while each datapoint is usually sampled uniformly in supervised learning. In this work, we focus mainly on *policy evaluation*, i.e evaluating the expected return of a fixed policy. More precisely, we study the behavior of Temporal Difference learning [8] which has become ubiquitous for policy evaluation in modern Deep RL [9, 10].

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

More precisely, our contributions are, in the limit of large number of states and parameters: **1)** We introduce a class of random graphs for which we can compute analytically the spectrum of the TD operator when using uniform off-policy sampling, **2)** we derive approximations for the largest and smallest eigenvalue in the non-uniform case. Furthermore, we analyze the behavior of the traditional Mean Square Bellman Error (MSBE) when learning with TD(0) and show that **3)** it exhibits a double descent phenomenon similar to supervised learning but with a multiplicative factor depending on the discount factor in the overparameterized regime and **4)** the asymptotic residual MSBE displays a peaking behavior near the interpolation threshold (when the number of parameters approaches the number of states) that is specific to RL. Finally, **5)** we showcase how the non-stationarity of the policy can make the optimization of the value function unstable. All of these points are motivated theoretically and empirically validated on small gridworlds MDPs.

2 Background

2.1 Value estimation

Definition: We place ourselves in the Markov Decision Process (MDP) setting [11] defined by $\{\mathcal{S}, \mathcal{A}, \mathbf{P}_{sa\to s}, \mathbf{r}, \gamma, p_0\}$, where $\mathcal{S}$ and $\mathcal{A}$ denote the finite state and action spaces, $\mathbf{P}_{sa\to s} \in \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|\times|\mathcal{A}|}$ is the environment transition probability matrix and $\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|}$ denotes the (deterministic) reward vector. $\gamma \in [0, 1[$ is the discount factor, and lastly, p_0 is the initial state distribution. For a given *policy* π, i.e a distribution over actions given a state, the goal of value estimation is to be able to estimate the expected discounted return $\mathbf{q}^\pi(s,a) = \mathbb{E}_{\mathbf{P}_{sa\to s},\pi}\big[\sum_t \gamma^t \mathbf{r}(s_t, a_t)|s_0 = s, a_0 = a\big]$. This quantity is also known as the state-action value or the Q-function. Furthermore, it can be written as the solution of the *Bellman equation* [12][1]: $\mathbf{q}^\pi(s,a) = \mathbf{r}(s,a) + \gamma\mathbb{E}_{\mathbf{P}_{sa\to s},\pi}[\mathbf{q}^\pi(s',a')]$. By defining $\mathbf{P}^\pi \in \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|\times|\mathcal{S}|\cdot|\mathcal{A}|}$ the state-action to state-action transition matrix, we can write it in matrix form

$$\mathbf{q}^\pi = \mathbf{r} + \gamma\mathbf{P}^\pi\mathbf{q}^\pi \tag{1}$$

Linear function approximation: While Equation (1) can be solve efficiently when $\mathbf{r}, \mathbf{P}^\pi$ are known and when $n = |\mathcal{S}| \cdot |\mathcal{A}|$ is not too large, in many cases this cannot be computed. One solution is to introduce a parameter vector $\theta \in \mathbb{R}^p$ and a feature matrix $\Phi \in \mathbb{R}^{n\times p}$ and optimize θ to approximate $\mathbf{q}^\pi$ as $\mathbf{q}^\pi \approx \Phi\theta$. This is referred to as the Linear Function Approximation (LFA) setting.

Temporal Difference with LFA: Estimating the Q-function in the Linear Function Approximation setting, can be done using the well-known algorithm TD(0) [8]

$$\theta_{t+1} = \theta_t + \eta\Phi^\top\Xi\big(\mathbf{r} + \gamma\mathbf{P}^\pi\Phi\theta_t - \Phi\theta_t\big) \tag{2}$$

where Ξ is the off-policy distribution matrix, a $n \times n$ diagonal matrix with positive entries summing to 1, weighting each state-action by their probability of being sampled. Contrary to gradient descent with an appropriate step-size, TD(0) is not guaranteed to converge as the matrix $\Phi^\top\Xi\big(\mathbf{I}_n - \gamma\mathbf{P}^\pi\big)\Phi$ might have a negative eigenvalue. This phenomenon is known as the *deadly triad* as it can only happen when we are (i) off-policy, (ii) using function approximation and (iii) bootstrapping on the next value as TD(0) does [13, 14, 15].

2.2 Random Matrix Theory

Wishart matrices: Extremely relevant in the study of least squares with random features are matrices from the *Wishart distribution* [16]. These matrices can be written as $\frac{1}{n}\Phi^\top\Phi$ for Φ a $n \times p$ matrix with entries of 0 mean, unit variance and bounded 4th moment sampled independently and identically distributed. In that case, when $p, n \to \infty$ and the ratio converges to a finite limit $\lim_{p,n\to\infty} p/n = \rho \in]0, +\infty[$, the empirical spectral distribution converges to the Marchenko-Pastur [17] (MP) distribution whose density can be written as

$$d\mu(x) = \max\{1 - 1/\rho, 0\}\ 1_{\{0\}}(x)dx + \frac{1}{2\pi\rho x}\sqrt{(x - \lambda^+_{\min})(\lambda_{\max} - x)}\ 1_{[\lambda^+_{\min},\lambda_{\max}]}(x)dx \tag{3}$$

[1]Note that we study here the expectation Bellman equation, not the optimality Bellman equation which would feature a $\max$ operator over the next action.

where $1_A(\cdot)$ is the indicator function, valued at 1 if $x \in A$, 0 otherwise and $\max\{1 - 1/\rho, 0\}\, \delta_{\{0\}}(x)$ are the 0 eigenvalues, which happens when $p \geq n$, $\lambda_{\max} = (1 + \sqrt{\rho})^2$, the largest, and $\lambda^+_{\min} = (1 - \sqrt{\rho})^2$, the smallest non-zero eigenvalue.

Wigner-type matrices: [18] historically introduced and studied symmetric $n \times n$ matrices of the form $\mathbf{W}$ where $\mathbf{W}_{ij}, i \geq j$ are sampled i.i.d from a law of mean 0 and unit variance for $i \geq j$[2]. $\frac{1}{\sqrt{n}}\mathbf{W}$ has a real spectrum which converges to the *semi-circle* distribution $d\mu(x) = \frac{1}{2\pi}\sqrt{4 - x^2}\, 1_{[-2,2]} dx$.

When non-symmetric, the spectrum converges to the complex unit disk$\{z \in \mathbb{C}, |z| \leq 1\}$[3] [20].

2.3 Random graphs

Random graphs are useful models to infer structural or spectral properties of typical graphs. The earliest theoretical analyses of random graphs that we are aware of were done independently by [21] and [22]. The model introduced in [21] is a random graph where each of the n nodes is connected at random to d others. The second model [22] is slightly different as each node is independently connected to any other with a probability $\mu \in]0, 1[$. We refer to the first one as the d-regular Erdős-Rényi model $\mathbb{G}(n, d)$ and the second one as the Erdős-Rényi-Gilbert $\mathbb{G}(n, \mu)$ model.

Given the adjacency matrix $\mathbf{A}$, with $\mathbf{A}_{ij} = 1$ if i is connected to j, it is possible to construct a matrix $\mathbf{P}$ by renormalizing the rows of $\mathbf{A}$ so that $\mathbf{P}$ is a stochastic matrix (positive entries and rows summing to 1). $\mathbf{P}$ corresponds to the Markov transition matrix of a random walk on the graph and is connected to the *(left)-normalized Laplacian* $\mathbf{L}$ by $\mathbf{L} = \mathbf{I}_n - \mathbf{P}$. For many classes of random graphs, the spectrum of these quantities is known [23, 24].

3 Spectrum of TD: from random graphs to spiked models

3.1 From Ordinary Least Squares to TD(0)

Random matrix models have been used in supervised learning, especially in the simple least squares model [25, 26, 27], to explain phenomena arising in deep learning. Indeed, while much simpler than large neural networks, these models can shed light on the optimization and generalization behavior of vastly overparameterized models [28, 29], something that most traditional complexity measures could not [30, 31].

Using gradient descent on a least square problem, our parameter update can be written as

$$\theta_{t+1} - \theta^* = \big(\mathbf{I}_p - \eta\Phi^\top\Phi\big)(\theta_t - \theta^*) \tag{4}$$

for $\Phi : \mathbb{R}^{n \times p}$ the feature matrix and θ^* the solution found by gradient descent. When choosing the learning rate η optimally, the convergence rate is expressed as $\frac{\lambda_{\max}/\lambda^+_{\min} - 1}{\lambda_{\max}/\lambda^+_{\min} + 1}$ for $\lambda^+_{\min}$, respectively $\lambda_{\max}$, the smallest, resp. largest, non-zero eigenvalue of $\frac{1}{n}\Phi^\top\Phi$ [32].

When Φ has i.i.d entries with 0 mean, unit variance and bounded 4th moment, as we will consider in the rest of this paper, $\frac{1}{n}\Phi^\top\Phi$ follows the Wishart distribution. In particular, we have analytical expressions for the asymptotic minimum and maximum eigenvalue (section 2.2). When $\lim_{p,n\to\infty} p/n = \rho$, we have $\lambda_{\max}/\lambda^+_{\min} \to \left(\frac{1+\sqrt{\rho}}{1-\sqrt{\rho}}\right)^2 \geq 1$. This function has asymptotes at 1, for $p = o(n)$ or $n = o(p)$ and diverges when $n = p$. Note that this peak-shaped function can also be observed empirically in more complex models such as kernel methods [33] and neural networks in the limit of infinite width as they are in the Neural Tangent Kernel regime [34, 35].

However, when using TD(0), presented in section 2.1, the feature covariance matrix $\Phi^\top\Phi$ would be replaced by $\Phi^\top\Xi(\mathbf{I}_n - \gamma\mathbf{P}^\pi)\Phi$. While we can model Φ as an i.i.d matrix, it is not obvious how to model Ξ and $\mathbf{P}^\pi$ in a realistic manner that would still allow us to characterize the spectrum of the

[2] and an additional light tail decay condition, weaker than the bounded 4-th moment one [19].

[3] both in probability and almost surely.

TD operator $\Xi(\mathbf{I}_n - \gamma \mathbf{P}^\pi)$. We believe this is one of the key reasons why overparameterization in Temporal Difference learning has not been studied extensively. Finding good typical models for $\mathbf{P}^\pi$ is therefore of utmost importance, and the next section will introduce a class of models for which the spectrum of TD can be computed.

3.2 $\mathbf{P}^\pi$ as the Markov transition matrix of a random graph

In RL, $\mathbf{P}^\pi$ is the Markov transition matrix of a random walk when sampling from π. As random graphs are used to model properties of typical graphs, a candidate for a model of $\mathbf{P}^\pi$ would be to take the Markov transition matrix of a random graph. However, random graphs, at least the Erdős-Rényi(-Gilbert) type ones, are not able natively to take into account the impact of π, most importantly the fact that some state-actions are visited more often than others under $\mathbf{d}_\pi$, the stationary distribution of π. In this subsection, we propose a simple model for $\mathbf{P}^\pi$, denoted by $\hat{\mathbf{P}}^\pi$, which is the Markov transition matrix of a $\mathbb{G}(n, \mu)$ graph *deformed* by $\mathbf{d}_\pi$.

Let us start with an undirected graph of the type $\mathbb{G}(n, \mu)$. To take into account the state-action visitation $\mathbf{d}_\pi$, we construct the diagonal matrix $\mathbf{D}_\pi$ so that $\mathbf{D}_\pi = \texttt{diag}(\mathbf{d}_\pi)$. Now, we define the deformed adjacency matrix $\mathbf{A}^\pi$ as $\mathbf{A}^\pi \triangleq n\mathbf{A}\mathbf{D}_\pi$ where $\mathbf{A}$ is the original adjacency matrix. Note here that when $\mathbf{d}_\pi$ is uniform, we revert back to the original case $\mathbf{A}^\pi = \mathbf{A}$. This construction keeps the same edges as in the original graph but re-weights them by how likely $\mathbf{d}_\pi$ would visit that state-action. We can informally write the Markov transition matrix associated to this deformed graph as[4] $\hat{\mathbf{P}}^\pi = \mathbf{1}_n \mathbf{d}_\pi^\top + \mathbf{X}^\pi$ where (i) $\mathbf{1}_n \mathbf{d}_\pi^\top$ is a deterministic rank one matrix associated to the eigenvalue 1 as $\mathbf{1}_n^\top \mathbf{d}_\pi = 1$ and (ii) $\mathbf{X}^\pi = \sqrt{\frac{n(1-\mu)}{\mu}} \frac{1}{\sqrt{n}} \mathbf{W}\mathbf{D}_\pi$ is a stochastic matrix of null expectation as $\frac{1}{\sqrt{n}}\mathbf{W}$ is a Wigner-type matrix with 0 mean and unit variance. It can be tempting to understand under which conditions the stochastic part becomes negligible and what it entails. As we will see shortly, this setting will reveal itself to be of importance when studying the spectrum of TD.

Definition 1 (Asymptotically well-connected graphs)**.** *We say a graph $\mathbb{G}$ is asymptotically well-connected if $|\lambda_2| \in o(1)$ for $1 = \lambda_1 \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$ the (modulus) ordered eigenvalues of its Markov transition matrix.*

The choice for the name of Definition 1 comes from the fact that $|\lambda_2|$ is related to the presence of *bottlenecks* in the graph through Cheeger's inequality [36].

Example. Graphs of the type $\mathbb{G}(n, \mu)$ (resp. $\mathbb{G}(n, d)$) are asymptotically well-connected if their expected degree $d = n\mu$ (resp. d) grows to $+\infty$ with n, i.e $1 \in o(d)$ [23, 24].

Proposition 3.1. *For the $\mathbf{d}_\pi$-deformed $\mathbb{G}(n, \mu)$ graph studied above, assuming $\hat{\mathbf{P}}^\pi = \mathbf{1}_n \mathbf{d}_\pi^\top + \mathbf{X}^\pi$, then the graph is asymptotically well-connected if $n\|\mathbf{d}_\pi\|_\infty \in o(\sqrt{d})$.*

Note here that this again generalizes the uniform case as, when $\mathbf{d}_\pi$ is uniform, the condition becomes equivalent to $1 \in o(d)$ as in the example. In the non-uniform case, this condition entails that either $\mathbf{d}_\pi$ spreads efficiently across states, or the degree d rises fast enough that those states are still visited sufficiently often. In that case, we have

$$\boxed{\hat{\mathbf{P}}^\pi = \mathbf{1}_n \mathbf{d}_\pi^\top + o(1)} \tag{5}$$

This approximation is particularly interesting as it asymptotically conserves the left and right eigenvectors, $\mathbf{d}_\pi$ and $\mathbf{1}_n$, of the original $\mathbf{P}^\pi$ associated to the eigenvalue 1. Furthermore, if all state-actions are reachable, the sequence of matrix power iterates of $\mathbf{P}^\pi$, $\left(\mathbf{P}^\pi\right)^k$ converges to $\mathbf{1}_n \mathbf{d}_\pi^\top$. While we are not the first to use tools from graph theory in the context of RL, related work we are aware of [37, 38, 39, 40] takes a different approach and leverages eigenvectors of the Markov transition matrix (or Laplacian) to construct better representations for the value function or use them as options. In contrast, we use spectral properties of $\mathbf{P}^\pi$ to study the convergence of Temporal Difference learning.

[4] Derivation can be found in Appendix.

3.3 Spiked Wishart model for TD with uniform sampling

Now that we have a simple random graph model for $\mathbf{P}^\pi$, we focus on estimating the spectrum of $\Phi^\top \Xi(\mathbf{I}_n - \gamma \hat{\mathbf{P}}^\pi)\Phi$. For simplicity, we assume the off-policy sampling is done uniformly, i.e $\Xi = \frac{1}{n}\mathbf{I}_n$. While this assumption is not realistic in most RL settings, it is still of interest in the context of dynamic programming where we have access to all the states.

Proposition 3.2 (Spiked MP)**.** *If $\rho\gamma^2 < 1$ the spectrum of $\frac{1}{n}\Phi^\top(\mathbf{I}_n - \hat{\mathbf{P}}^\pi)\Phi$ converges to the Marchenko-Pastur law with parameter ρ. When $\rho\gamma^2 \geq 1$, there is a phase transition [41] for the minimum non-zero eigenvalue $\lambda^+_{\min}$ which separates from bulk and converges to*

$$\boxed{\lambda^+_{\min} \xrightarrow[p/n \to \rho]{a.s} \lambda^{spiked} = (1-\gamma)(\rho - \tfrac{1}{\gamma})} \tag{6}$$

Note that the phase transition is smooth as for $\rho = 1/\gamma^2$, $\lambda^{spiked} = (\sqrt{1/\gamma^2} - 1)^2$.

Vast overparameterization and tabular case: An interesting regime is when $\rho \gg \frac{1}{\gamma^2} > 1$. In that case, the maximum eigenvalue of the MP law and the spiked one differ by a factor $H = \frac{1}{1-\gamma}$, the *effective horizon* of the problem, as $\lim_{\rho\to\infty} \lambda_{\max}/\lambda^+_{\min} = \frac{1}{1-\gamma}$. In the tabular regime, $\Phi = \mathbf{I}_n$. As $\mathbf{P}^\pi$ has eigenvalues in $[-1, 1]$, the eigenvalues of $\mathbf{I}_n - \gamma\mathbf{P}^\pi$ are in $[1-\gamma, 1+\gamma]$. This would give a worst case ratio of $\frac{1+\gamma}{1-\gamma}$ which, up to a factor $1+\gamma \leq 2$, is the same as the vastly overparameterized case above. As these eigenvalues govern the optimization speed of TD, the vastly overparameterized case and the tabular can be thought of as comparable.

3.4 Non-uniform off-policy sampling

In the previous subsection we were able to characterize the spectrum of $\Phi^\top \Xi(\mathbf{I}_n - \gamma\hat{\mathbf{P}}^\pi)\Phi$ assuming $\Xi = \frac{1}{n}\mathbf{I}_n$. In this subsection we analyze the more general case where Ξ is not uniform. However by doing so, we lose guarantees and must resort to approximations. First, we analyze the eigenvalues of $\Xi(\mathbf{I}_n - \gamma\hat{\mathbf{P}}^\pi)$, or rather, in the limit of large n, the eigenvalues of $\Xi(\mathbf{I}_n - \gamma\mathbf{1}_n\mathbf{d}_\pi^\top)$.

Lemma 3.1. *If $\Xi = \texttt{diag}(\xi_1, \ldots, \xi_n)$ is non-singular, the eigenvalues of $\Xi(\mathbf{I}_n - \gamma\mathbf{1}_n\mathbf{d}_\pi^\top)$ satisfy a* secular *equation [42]*

$$1 - \gamma \sum_{i=1}^{n} \frac{\xi_i \cdot \mathbf{d}_\pi(i)}{\xi_i - \lambda} = 0 \tag{7}$$

This expression is another form of the *characteristic polynomial* of a matrix whose roots are its eigenvalues. However, this form enables us to derive approximations more easily. In particular, for the least visited state-action $m = \arg\min_{i=1,\ldots,n} \xi_i$, if $\xi_m \ll \xi_i, i \neq m$, we have for $\lambda \approx \xi_m$ that $\frac{1}{\xi_i - \lambda} \approx \frac{1}{\xi_i}$. From this assumption, we derive an approximation for the minimum eigenvalue of $\Xi(\mathbf{I}_n - \gamma\hat{\mathbf{P}}^\pi)$

$$\ell_{\min} \approx \xi_m \tfrac{1-\gamma}{1-\gamma+\gamma\,\mathbf{d}_\pi(m)} \tag{8}$$

This eigenvalue is inferior to ξ_m and reflects the interaction between the off-policy distribution and the discount factor. At worst, if $\mathbf{d}_\pi$ mostly visits state m, it will be close to $\xi_m(1-\gamma)$ which is the product of the minimum eigenvalues of Ξ and $\mathbf{I}_n - \gamma\hat{\mathbf{P}}^\pi$. Note that a similar expression can be derived for the maximum eigenvalue under similar assumptions.

Now, again reasoning approximately, if that eigenvalue was very small compared to the other ones who would form a bulk of similar values, we could expect $\Phi^\top \Xi(\mathbf{I}_n - \gamma\hat{\mathbf{P}}^\pi)\Phi$ to have a similar spiking behavior as before. However, in that case, the spiking eigenvalue would be

$$\lambda^+_{\min} \approx n\ell_{\min}\big(\rho - \tfrac{1}{1-n\ell_{\min}}\big) \tag{9}$$

While less simple than the estimators developed in the last subsection, we will see in Section 5 that it still predicts accurately the smallest non-zero eigenvalue.

4 Optimizing with Temporal Difference learning

4.1 Decomposition of the error

Using results about the spectrum of the TD operator shown in the last section, we analyze the behavior of the Mean Square Bellman Error $\mathcal{L}_{MSBE}(\mathbf{q}_t) = \frac{1}{2}\|\mathbf{r} + \gamma \mathbf{P}^\pi \mathbf{q}_t - \mathbf{q}_t\|_\Xi^2$ during the optimization. In particular, assuming $\Phi^\top \Xi(\mathbf{I}_n - \gamma \mathbf{P}^\pi)\Phi$ is non-singular for $p < n$

$$\mathbf{q}_{\text{TD}} = \begin{cases} \Phi\theta_{\text{TD}} = \Phi(\Phi^\top \Xi(\mathbf{I}_n - \gamma \mathbf{P}^\pi)\Phi)^{-1}\Phi^\top \Xi \mathbf{r}, \text{ if } p < n \\ \mathbf{q}^\pi \quad = \quad (\mathbf{I}_n - \gamma \mathbf{P}^\pi)^{-1}\mathbf{r}, \text{ if } p > n \end{cases}$$

the $\mathbf{q}$-value asymptotically reached by TD(0), we can decompose this error into two components.

Lemma 4.1 (Decomposition of the error)**.** *For $\mathbf{q}_t = \Phi\theta_t$, where θ_t is updated with TD(0) (eq.* (2)*), we have the following decomposition*

$$\|\mathbf{r} + \gamma \mathbf{P}^\pi \mathbf{q}_t - \mathbf{q}_t\|_\Xi \leq \sqrt{n\|\Xi\|_\infty} \cdot \|\mathbf{q}_t - \mathbf{q}_{TD}\|_\Xi + \|\mathbf{r} + \gamma \mathbf{P}^\pi \mathbf{q}_{TD} - \mathbf{q}_{TD}\|_\Xi \tag{10}$$

where $\|\mathbf{x}\|_\Xi = \sqrt{\mathbf{x}^\top \Xi \mathbf{x}}$ is the - norm in metric Ξ.

Following [43], we refer to $\|\mathbf{q}_t - \mathbf{q}_{\text{TD}}\|_\Xi$ as the approximation error and $\|\mathbf{r} + \gamma \mathbf{P}^\pi \mathbf{q}_{\text{TD}} - \mathbf{q}_{\text{TD}}\|_\Xi$ as the estimation error. The factor $\sqrt{n\|\Xi\|_\infty} \geq 1$ comes from bounding the spectral norm of $(\mathbf{I}_n - \gamma \mathbf{P}^\pi)$ in metric Ξ. Note that, beyond this factor, we are not exactly in the same setting as [43] as the asymptotic TD(0) value estimate $\mathbf{q}_{\text{TD}}$ may not be the one reaching the lowest estimation error, as known in the RL community [44]. Indeed, TD(0) finds the minimum of the Mean Square Projected Bellman Error (MSPBE) but not of the MSBE [45]. However in practice it is common to minimize the MSBE as it does not require an expensive projection step.

4.2 The ordinary behavior of the approximation error

In this subsection, we take a look at the approximation error, i.e how efficiently we converge to $\mathbf{q}_{\text{TD}}$.

Proposition 4.1. *Let us assume that $\Phi^\top \Xi(\mathbf{I}_n - \gamma \mathbf{P}^\pi)\Phi$ is diagonalizable as $\mathbf{Q}\Lambda\mathbf{Q}^{-1}$, and that its spectrum is real and positive. Denoting by $\lambda_{\min}^+$ and $\lambda_{\max}$ its smallest non-zero and largest eigenvalues, for $\eta = \frac{2}{\lambda_{\min}^+ + \lambda_{\max}}$*

$$\|\mathbf{q}_t - \mathbf{q}_{TD}\|_2^2 \leq \left(\frac{\lambda_{\max}/\lambda_{\min}^+ - 1}{\lambda_{\max}/\lambda_{\min}^+ + 1}\right)^{2t} K^2 \|\mathbf{q}_0 - \mathbf{q}_{TD}\|_2^2 \tag{11}$$

where $K = \kappa(\Phi)\kappa(\mathbf{Q})$ when $p < n$ and $K = \kappa(\mathbf{Q})$ when $p > n$ where κ is the condition number.

The assumptions above can seem strong however they are verified under our model (cf Section 3.3). This is a slightly more general version of the linear convergence of gradient descent, however, as the TD operator may not be diagonalizable in an orthonormal basis, we incur an additional factor K^2.

4.3 The singular behavior of the estimation error

Now we turn our attention to the second term; the estimation error which is the *residual error* after $T \to +\infty$ steps. In supervised learning, we expect this error to be the usual bias of our algorithm and thus to decrease as p/n increases until $p = n$ for which the bias would reach 0 as it is possible to interpolate the data. However, in the reinforcement learning, the behavior of this term can be quite different, in particular it can be non-monotonous.

Proposition 4.2. *Let us assume $\hat{\mathbf{P}}^\pi$ satisfies the well-connected property, i.e that it is asymptotically rank one*

$$\lim_{\substack{n,p\to\infty \\ p/n \to \rho}} \|(\mathbf{I}_n - \gamma\hat{\mathbf{P}}^\pi)\mathbf{q}_{TD} - \mathbf{r}\|_\Xi^2 = \mathbf{r}^\top \Pi_\perp \mathbf{r} + \frac{2\gamma}{1-\gamma\tilde{\rho}}(\mathbf{d}_\pi^\top \Pi \mathbf{r})(\xi^\top \Pi_\perp \mathbf{r}) + (\frac{\gamma}{1-\gamma\tilde{\rho}})^2 \xi^\top \Pi_\perp \mathbf{1}_n (\mathbf{r}^\top \Pi \mathbf{d}_\pi)^2$$

where Π is the projector onto the span of Φ in norm Ξ: $\Pi = \Phi(\Phi^\top \Xi \Phi)^{-1}\Phi^\top \Xi$, $\Pi_\perp = \mathbf{I}_n - \Pi$ is its complement, and $\tilde{\rho} = \mathbf{1}_n^\top \Pi \mathbf{d}_\pi$.

When $\gamma = 0$, i.e we just perform reward regression, this is equal to $\mathbf{r}^\top \Pi_\perp \mathbf{r}$ which is exactly the residual error in $\mathbf{r}$ that our representation cannot capture. When $\rho > 1$, $\Pi_\perp = 0$ so the limit is zero as we expect from a model able to interpolate the data. However, there is a behavior here not present in supervised learning. When $\tilde{\rho} \approx 1$, which we can expect when Π is closer to identity, i.e p close to but less than n, a *peaking* behavior of the estimation error can happen.

Replacing Π by its expectation[5] $\rho \mathbf{I}_n$ above can lead to an equation simpler to analyze, albeit biased

$$1_{]0,1]}(\rho)\ (1-\rho)\big(\|\mathbf{r}\|_\Xi^2 + 2\frac{\gamma\rho}{1-\gamma\rho}(\pi^\top \mathbf{r})(\mathbf{d}^\top \mathbf{r}) + (\pi^\top \mathbf{r})^2(\frac{\gamma\rho}{1-\gamma\rho})^2\big)$$

This phenomenon might appear counter-intuitive as we would expected our error to decrease with p. We provide a few insights to make sense of this. First, when γ increases, so does the scale of the Q-function, so we can expect our errors to get bigger. Second, using random features Φ informally has a *smoothing* effect, stronger when ρ is small, which causes the reduction in magnitude from $\frac{1}{1-\gamma}$ to $\frac{1}{1-\gamma\rho}$. This can be observed in the proof in appendix. Finally, peaking behaviors of TD such as this one have been observed in the community [44] but not theoretically explained in a general setting as far as we know.

4.4 Beyond stationary sampling

Until now, we always mentioned n as the number of states with the assumption that Ξ was non-singular, i.e that our off-policy sampling *visited all the states*. Here, we take a closer look at what happens when this is not the case. In particular, we study a simple toy model where we assume Ξ only visits $n_{\text{vis}}(t)$ states at time t with uniform probability. This simply changes all the n by $n_{\text{vis}}(t)$ in the previous results. As often observed during training, at $t = 0$, π might start as a more exploratory policy ($n_{\text{vis}}(t = 0)$ high) and over time converges on a more deterministic solution ($n_{\text{vis}}(t)$ low). This can be problematic when performing TD learning with a fixed stepsize η. In that case, under the spiked model developed in Section 3.3 we need to ensure $\eta \leq {}^2/\big(1+\sqrt{p/n_{\text{vis}}(t)}\big)^2$ or equivalently

$$n_{\text{vis}}(t) \geq {}^p/\big(\sqrt{\tfrac{2}{\eta}}-1\big)^2 \tag{12}$$

As $n_{\text{vis}}(t)$ can typically decrease, **this condition can be broken and thus TD learning will diverge**. We conjecture this phenomenon may related to the popularity of adaptative step size methods in RL over fixed step size ones.

5 Experiments

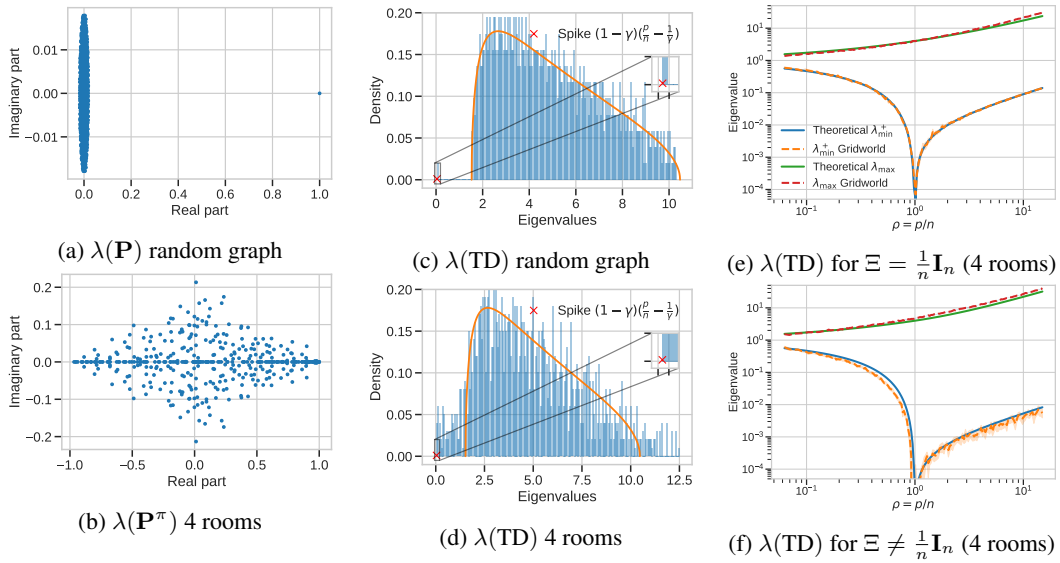
In all the experiments we use Φ i.i.d with normal entries. Per the universality property, our theoretical results hold for all distributions satisfying the hypotheses in Section 2.2.

Study of the spectral distribution: On Figure 1, we validate the theory and approximations of Section 3. For an asymmetric random graph $n = 5000, \mu = {}^1/_2$, the spectrum of $\mathbf{P}$ (a) has an eigenvalue at 1 and the rest of order $\frac{1}{\sqrt{n}}$ [47]. As predicted, the spectrum of the TD operator matches closely a spiked MP distribution (b). When the graph does not satisfy the well-connected property, for instance on a 4 rooms domain ($n = 1600$, π stochastic[6]) (c) the spectral distribution is more complex (d). Despite this, we show that the empirical and theoretical (eq. (9)) values for $\lambda_{\min}^+$ and $\lambda_{\max}$ on the 4 rooms domain ($n = 256$) closely match, both for Ξ uniform or highly ill-conditioned with evenly spaced ξ_i so that $\frac{\max_i \xi_i}{\min_i \xi_i} = 2000$. While (b) and (d) show how violating the *asymptotically well-connected* property breaks our spiked MP distribution result, we show on (e) and (f) that the extreme eigenvalues, which are the ones that matter for our optimization results, are still well predicted, even in more complex settings.

Optimization behavior of TD: On Figure 2 we discuss the results of Section 4. We use again a 4 rooms domain with a fixed stochastic policy and study the Bellman error during training. On

[5] when $\Xi = \frac{1}{n}\mathbf{I}_n$, by rotational invariance of Φ, we $\mathbb{E}[\Pi] = \rho \mathbf{I}_n$, thus $\mathbb{E}[\tilde{\rho}] = \rho$.

[6] We sampled $\pi(a|s) \sim \text{Uniform}[0, 1]$ and renormalized.



(a) $\lambda(\mathbf{P})$ random graph

(b) $\lambda(\mathbf{P}^\pi)$ 4 rooms

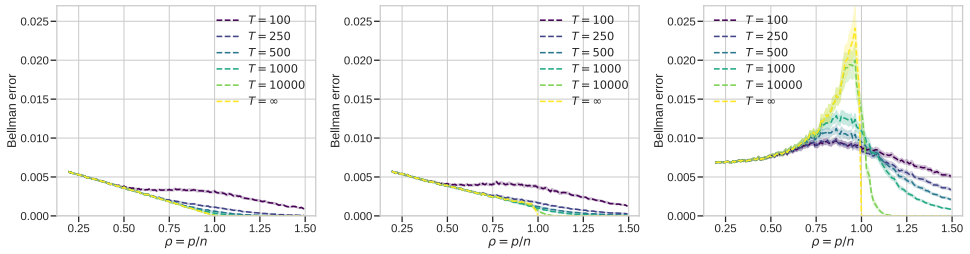
(c) $\lambda(\text{TD})$ random graph

(d) $\lambda(\text{TD})$ 4 rooms

(e) $\lambda(\text{TD})$ for $\Xi = \frac{1}{n}\mathbf{I}_n$ (4 rooms)

(f) $\lambda(\text{TD})$ for $\Xi \neq \frac{1}{n}\mathbf{I}_n$ (4 rooms)

Figure 1: (a) and (b): scatterplot of the spectra of the Markov transition matrix for a $\mathbb{G}(n, \mu)$ graph and the 4 rooms domain. (c) and (d): spectral distribution for the TD operator for $\rho = 5$. When the well-connected assumption is verified (a), the spectral distribution is a spiked ($\times$) Marchenko-Pastur (MP) law (in orange) (c). When the assumption is violated and $\hat{\mathbf{P}}^\pi \not\approx \mathbf{1}_n \mathbf{d}_\pi^\top$ (b), the distribution bleeds out of the MP support, yet the minimum eigenvalue is still close to the spike (d). On (e) and (f) we compare our prediction of the extreme eigenvalues $\lambda^+_{\min}$ and $\lambda_{\max}$ of the TD operator on the 4 rooms domain [46] (b) as a function of p/n. We plot the empirical extreme eigenvalues (dotted) vs (e) MP theoretical ones (eq. (3)) for Ξ uniform and (f) approximation in eq. (9) for Ξ non-uniform. Shaded areas are a 95% confidence interval when randomizing Φ and π (10 seeds).



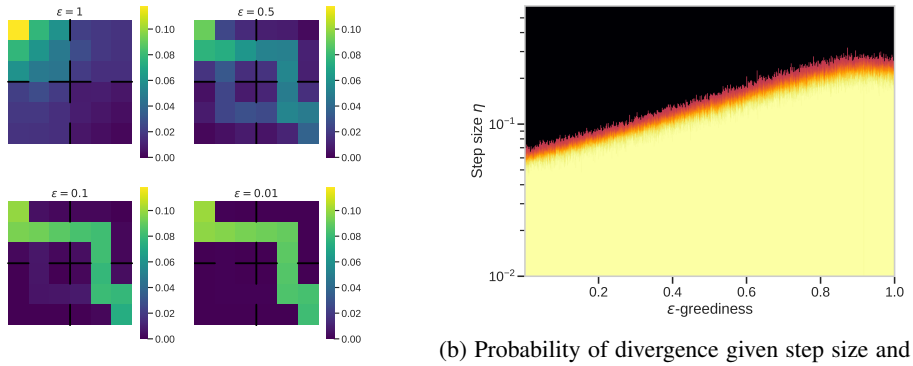
(a) $\gamma = 0.1$, $\mathbf{d}_\pi^\top \mathbf{r} \approx 0.08$

(b) $\gamma = 0.98$, $\mathbf{d}_\pi^\top \mathbf{r} \approx 0.02$

(c) $\gamma = 0.98$, $\mathbf{d}_\pi^\top \mathbf{r} \approx 0.07$

Figure 2: Value of the MSBE as a function of the ratio p/n for various iterations T during optimization. $T = +\infty$ corresponds to the asymptotic solution found by TD(0) (eq. (10)). We use uniform sampling ($\Xi = \frac{1}{n}\mathbf{I}_n$). On (a), for low γ we observe a similar to supervised learning where the residual error decreases and reaches 0 at $p = n$. (b) displays a similar behavior as the term $\mathbf{d}_\pi^\top \mathbf{r}$ is small. (c) As $\mathbf{d}_\pi^\top \mathbf{r}$ is higher, the *peaking* terms $\frac{1}{1-\gamma\tilde{\rho}}$ are non-negligible. Shaded areas are a 99% confidence interval over randomness of θ_0 and Φ (100 seeds).

all subfigures of Figure 2, we observe that for $\rho \gg 1$ or $\rho \ll 1$, the optimization is faster as the curves for low T are much closer to the asymptotic solution $T = +\infty$ (computed analytically) than for $\rho \approx 1$. This is in line with the result of section 4.2 for a spike model: the ratio of eigenvalues which governs the optimization speed is $\lambda_{\max}/\lambda^+_{\min} \to 1$ for $\rho \ll 1$ and to $\frac{1}{1-\gamma}$ for $\rho \gg 1$ while it diverges for $\rho \to 1$. This leads to a classical *double descent* phenomenon on (a) and (b) purely due to optimization as predicted by [7]. On (a) and (b) because either γ or $\mathbf{d}_\pi^\top \mathbf{r}$ is small, the terms of order $\frac{1}{1-\gamma\tilde{\rho}}$ are negligible and thus it behaves similarly to supervised learning. On (c), the term $\frac{1}{1-\gamma\tilde{\rho}}$ dominates and we observe the peaking behavior of the estimation error (Section 4.3). Furthermore



(a) $\mathbf{d}_\pi$ on 4 rooms for different ϵ.

(b) Probability of divergence given step size and ϵ-greediness.

Figure 3: (a): Example of the discounted state distribution in a small 4 rooms domain for varying ϵ-greedy policies. (b) The probability of divergence of TD as a function of the step size η and the ϵ-greediness. For each point we report the probability it leads to divergence: from yellow (0%) to black (100%).

on Figure 4 in appendix we study how our model and its approximation of Section 3.3 match the observed peaking behavior.

Stability and non-stationarity: On Figure 3 we highlight how Ξ can affect the convergence of TD ($p = 200$). For this, we use a small 4 rooms ($n = 144$) environment[7] and compute an optimal deterministic policy π^*. Then, by varying the ϵ of a ϵ-greedy policy based on π^*, we can analyze which combination of step size η and ϵ would lead to convergence. On (a), we show the discounted state distribution for various ϵ where larger ϵ lead to more exploration. While our theoretical argument in Section 4.4 was about the number of visited states, we take here a more general and realistic approach where the state visitation is controlled by the greediness of our policy. Typically, during the optimization of a value based method, ϵ is annealed to 0 and the policy becomes more deterministic. On (b) we show probability of diverging (averaged over 50 samples of Φ) measured by computing the spectral radius of our iteration matrix. As predicted, for a fixed η, when ϵ decreases, we might go through a phase transition where the TD iteration, stable at the beginning, becomes unstable.

6 Conclusion

In this work, we have analyzed the spectral distribution of the Temporal Difference operator when using random features and modeling the Markov transition matrix as the one of a random graph. This characterization allowed us to make predictive theories for the optimization error and stability of TD. Notably we are able to predict the how fast we converge and whether the solution found is accurate as a function of the ratio between the number of parameters and the number of state-action pairs. In particular, we highlighted and theoretically explained several phenomena specific to reinforcement learning such as the peaking behavior of the estimation error of TD or how changing the policy might cause divergence when using a fixed step, even in the *on-policy* setting.

We believe our work paves to way toward an average-case theory of reinforcement learning algorithms in the same way this was done for supervised learning and used to understand deep learning.

Important directions for future work include: (i) improving the model for $\mathbf{P}^\pi$ through careful analysis of more expressive random graphs, (ii) derive non-asymptotic results using the fluctuations of the extreme eigenvalues of the MP law [41] (iii) extend our results to neural networks in the Neural Tangent Kernel regime and finally (iv) study the behavior of policy optimization algorithms.

[7] where the bottom right state with a reward of 1 leads back to the initial state on the top left.

References

[1] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

[2] Stuart Geman, Elie Bienenstock, and René Doursat. Neural Networks and the Bias/Variance Dilemma. *Neural Computation*, 4(1):1–58, 01 1992.

[3] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv: Arxiv-1810.04805*, 2018.

[6] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

[7] Ilja Kuzborskij, Csaba Szepesvári, Omar Rivasplata, Amal Rannen-Triki, and Razvan Pascanu. On the role of optimization in double descent: A least squares study. *Advances in Neural Information Processing Systems*, 34, 2021.

[8] Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.

[9] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[10] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018.

[11] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[12] Richard Bellman. The theory of dynamic programming. *Bulletin of the American Mathematical Society*, 60(6):503–515, 1954.

[13] Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning Proceedings 1995*, pages 30–37. Elsevier, 1995.

[14] John N Tsitsiklis and Benjamin Van Roy. Analysis of temporal-diffference learning with function approximation. In *Advances in neural information processing systems*, pages 1075–1081, 1997.

[15] Hado Van Hasselt, Yotam Doron, Florian Strub, Matteo Hessel, Nicolas Sonnerat, and Joseph Modayil. Deep reinforcement learning and the deadly triad. *arXiv preprint arXiv:1812.02648*, 2018.

[16] John Wishart. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, pages 32–52, 1928.

[17] Vladimir A Marčenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967.

[18] Eugene P. Wigner. Characteristic vectors of bordered matrices with infinite dimensions. *Annals of Mathematics*, 62(3):548–564, 1955.

[19] Terence Tao and Van Vu. Random matrices: The universality phenomenon for wigner ensembles. *arXiv preprint arXiv: Arxiv-1202.0068*, 2012.

[20] Terence Tao, Van Vu, and Manjunath Krishnapur. Random matrices: Universality of esds and the circular law. *The Annals of Probability*, 38(5):2023–2065, 2010.

[21] Paul Erdős, Alfréd Rényi, et al. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

[22] Edgar N Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144, 1959.

[23] Yufei Zhao. Spectral distributions of random graphs. 2012.

[24] Linh Tran, Van Vu, and Ke Wang. Sparse random graphs: Eigenvalues and eigenvectors. *arXiv preprint arXiv: Arxiv-1011.6646*, 2010.

[25] Yann Le Cun, Ido Kanter, and Sara A. Solla. Eigenvalues of covariance matrices: Application to neural-network learning. *Phys. Rev. Lett.*, 66:2396–2399, May 1991.

[26] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

[27] Michal Derezinski, Feynman T Liang, and Michael W Mahoney. Exact expressions for double descent and implicit regularization via surrogate random design. *Advances in neural information processing systems*, 33:5152–5164, 2020.

[28] Jeffrey Pennington and Yasaman Bahri. Geometry of neural network loss surfaces via random matrix theory. In *International Conference on Machine Learning*, pages 2798–2806. PMLR, 2017.

[29] Alexander Wei, Wei Hu, and Jacob Steinhardt. More than a toy: Random matrix models predict how real-world neural representations generalize. *arXiv preprint arXiv: Arxiv-2203.06176*, 2022.

[30] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

[31] Gintare Karolina Dziugaite, Alexandre Drouin, Brady Neal, Nitarshan Rajkumar, Ethan Caballero, Linbo Wang, Ioannis Mitliagkas, and Daniel M Roy. In search of robust measures of generalization. *Advances in Neural Information Processing Systems*, 33:11723–11733, 2020.

[32] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

[33] Tomaso Poggio, Gil Kur, and Andrzej Banburski. Double descent in the condition number. *arXiv preprint arXiv:1912.06190*, 2019.

[34] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. *arXiv preprint arXiv:1711.00165*, 2017.

[35] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

[36] Jeff Cheeger. A lower bound for the smallest eigenvalue of the laplacian. *Problems in analysis*, Princeton Univ. Press, Princeton, N. J.,:pp. 195–199., 1970.

[37] Sridhar Mahadevan and Mauro Maggioni. Proto-value functions: A laplacian framework for learning representation and control in markov decision processes. *J. Mach. Learn. Res.*, 8:2169–2231, 2007.

[38] Marlos C. Machado, Marc G. Bellemare, and Michael Bowling. A Laplacian framework for option discovery in reinforcement learning. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2295–2304. PMLR, 06–11 Aug 2017.

[39] Marlos C. Machado, Clemens Rosenbaum, Xiaoxiao Guo, Miao Liu, Gerald Tesauro, and Murray Campbell. Eigenoption discovery through the deep successor representation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[40] Yifan Wu, George Tucker, and Ofir Nachum. The laplacian in rl: Learning representations with efficient approximations. *arXiv preprint arXiv:1810.04586*, 2018.

[41] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

[42] Gene H Golub. Some modified matrix eigenvalue problems. *SIAM review*, 15(2):318–334, 1973.

[43] Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

[44] Bruno Scherrer. Should one compute the temporal difference fix point or minimize the bellman residual? the unified oblique projection view. *arXiv preprint arXiv:1011.4362*, 2010.

[45] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[46] Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

[47] Charles Bordenave, Pietro Caputo, and Djalil Chafaï. Circular law theorem for random markov matrices. *Probability Theory and Related Fields*, 152(3):751–779, 2012.