Sufficient Conditions for Divergence in Projected Bellman Equation Methods

Samuel Yang-Zhao

November 2018

A thesis submitted for the degree of Bachelor of Mathematical Sciences (Honours) of the Australian National University

Australian National University

For Guang, Eric and Huawei

Declaration

The work in this thesis is my own except where otherwise stated.

Samuel Yang-Zhao

Acknowledgements

A year and a half ago, I would have never thought that I would be just about to finish an honours thesis in mathematics. I am deeply indebted to the many people who have allowed me to come this far.

I would like to express my sincerest gratitude to my supervisors Marcus Hutter and Markus Hegland. To Marcus Hutter, thank you for providing me with this opportunity to undertake such a challenging yet interesting topic and for providing me the compass to eventually reach the destination. To Markus Hegland, thank you for being so generous with your time in our numerous discussions and helping me develop much more mathematical rigour. My appreciation for mathematics has never been higher.

Immeasurable thanks must go to Sultan. I will never forget the amount of time you spent proofreading my thesis and discussing my half-formed, and often nonsensical, ideas on a weekly basis. Thank you for being so patient when I randomly appear to ask you questions and for steering me in the right direction on numerous occasions.

I must thank my fellow honours cohort, especially all who were situated in the bullpen on level 3. To Feng, Chris and Kyle; the camaraderie with you boys is special. To Jane, thank you for being an outlier who never needs to sleep so that I always had at least one companion at the office. Thanks must also go to Omar; thank you for proofreading my thesis and the numerous discussions we had were immensely useful.

Penultimately, I must thank my girlfriend Michelle. Spending time with you this year has been a blessing. Thank you for keeping me sane.

Lastly, I would be remiss not to thank my parents. To mum, thank you for always being so caring and understanding. To dad, thank you for always providing timely advice. Without you guys none of this would be possible.

Contents

Notation and terminology

In the following, let $\mathcal{H}$ and $\mathcal{W}$ be two vector spaces and $A : \mathcal{H} \to \mathcal{W}$ a linear transformation.

Notation

$\mathbb{R}$	The set of real numbers.
$\mathbb{N}$	The set of natural numbers.
$\mathbb{Z}$	The set of integer numbers.
$\text{im}(A)$	The image of A, that is the set $\text{im}(A) \coloneqq \{Ax : x \in \mathcal{H}\}$.
$\ker(A)$	The kernel of A, that is the set $\ker(A) \coloneqq \{x \in \mathcal{H} : Ax = 0\}$.
$\text{span}(B)$	The span of B.
$\dim(\mathcal{H})$	The dimension of $\mathcal{H}$.
$x \perp y$	x is perpendicular to y.
$x \perp_\mu y$	Suppose we have an inner product $\langle \cdot, \cdot \rangle_\mu$. Then this notation denots that x is perpendicular to y with respect to $\langle \cdot, \cdot \rangle_\mu$.
$\mathcal{H} \oplus \mathcal{W}$	The direct sum of $\mathcal{H}$ and $\mathcal{W}$.
$\mu[\mathbb{R}]$	$\mu[\mathbb{R}] \coloneqq \int_{\mathbb{R}} \mu(s)ds$.
$\mathbb{E}(X)$	Let X be a random variable. $\mathbb{E}(X)$ denotes the expectation of X.
$\mathbf{1}_B$	The characteristic function for the set B.
$\mathbf{1}(s' = s)$	The indicator function.

Introduction

Reinforcement learning refers to both the problem of finding an optimal control policy for an agent to follow in an environment and also its various solution methods [17]. Typically the goal is to find a function, known as the value function, that generates the value of each state in the environment's state space. A control policy is then defined according to this value function. One of the most well-researched approaches of this form to date has been to resolve reinforcement learning problems using dynamic programming techniques. In particular, the task of finding the value function can be modelled as a fixed point problem with Bellman's equations. However, tradition dynamic programming techniques are often computationally intractable when the state space becomes large. This is the case in most scenarios of practical interest.

To resolve this issue, function approximation methods and sampling techniques are introduced to produce approximate solutions [2]. In fact, the most effective reinforcement learning algorithms utilise a combination of both elements. One of the simplest forms of function approximation is linear function approximation. In this case, approximate solutions are found in a lower-dimensional subspace and represented as a linear combination of basis functions. However, when using linear function approximation, the fixed point of the Bellman equation may no longer be an element of the lower-dimensional subspace as the Bellman operator that characterises the Bellman equation is an affine linear operator. Instead, many methods look to solve for an approximate solution by taking a projection of the Bellman equations [2].

A pertinent question to consider is under what conditions will reinforcement learning methods with linear function approximation converge ti the right solution? This traditional problem has been tackled quite extensively. For example, Tsitsiklas and Van Roy [21] were able to show that the $TD(\lambda)$ algorithm converges under conditions of ergodicity in the underlying state Markov chain and Baird [1] was able to show that the Residual Gradient algorithm converges ro-

bustly. Famous counter-examples also exist showing that $TD(0)$ diverges when the conditions derived by Tsitsiklis and Van Roy are violated. Recently, Sutton [17] detailed a subtle form of divergence that can occur for Bellman error based methods. In the example, Sutton shows that Bellman error based methods, of which the Residual Gradient method is one, may converge but to the *wrong solution*. Naturally, this leads to the question: when can this more subtle form of divergence occur?

In this thesis, that is the question that we look to investigate. In particular, we consider a class of algorithms whose solutions are based on calculating the solution of a projected Bellman equation and look to show sufficient conditions under which they diverge. Ultimately, we are able to characterise the results in the case of either a finite or continuous state space.

Before we proceed, it must be acknowledged that the ideas behind the results derived were summarised in [11]. Further reference to [11] will not be made for brevity. We now outline the content presented in this thesis.

0.1 Thesis Outline

In chapter 1 we present the preliminary mathematical concepts that are relied upon throughout this thesis. In particular, the emphasis is placed on results that pertain to stochastic processes, especially Markov processes, and Hilbert space theory.

In chapter 2, we introduce readers to the reinforcement learning problem and its underlying mathematical foundations. To begin with, we describe the agent-environment framework which encapsulates the general set-up in reinforcement learning. We then introduce a special type of environment for the reinforcement learning problem known as Markov decision processes. Markov decision processes are focused on throughout this thesis as most of the results in the literature are specified for this case. The mathematical foundations underpinning dynamic programming are then discussed to introduce readers to the main framework underpinning how reinforcement learning problems are solved.

In chapter 3 we introduce reader's to projected equation methods and set up the main class of algorithms that we consider divergence results for. To facilitate this, oblique projection operators are discussed before we focus in on a specific subset of such operators that are useful for finding computationally tractable approximate solutions. We call this set of projection operators the natural pro-

jection operators. Once this underlying theory is set-up, we turn our attention to characterising approximate solutions to the projected Bellman equations. When considering projected Bellman equations, we present a novel perspective on the key quantities that uniquely determine the solutions to a projected Bellman equation. This characterisation motivates us to define the set of natural algorithms upon which our results hold.

In chapter 4, we restrict our attention to the case of a finite state space and present a survey of well-known reinforcement learning methods. The purpose of this chapter is to show that our general set-up for natural algorithms also holds in the case of a finite state space. In particular, we are able to show that many well-known reinforcement learning algorithms do fall in our class of natural algorithms, thus justifying our class as a valid and interesting set to consider divergence results for.

In chapter 5, we present the main results of this thesis which are collectively known as the ambiguity conditions. We begin by providing an example that demonstrates the subtle type of divergence that can be experienced by a natural algorithm. We define this type of divergence as ambiguity and formalise it precisely. Once ambiguity is defined, we derive sufficient conditions under which ambiguity holds. Results are shown separately for the cases of a finite state space and a continuous state space. Our main result in the continuous case shows that there exists conditions under which *any* natural algorithm will either diverge or converge to the wrong solution.

Finally, in chapter 6 we consider extensions and open problems that arise as a consequence of the new results found.

Chapter 1

Preliminaries

In this chapter we a brief and succinct tour through the mathematical concepts and frameworks that will be utilised throughout. We assume a working knowledge of measure-theoretic probability theory.

1.1 Background

Throughout this thesis, we focus only on real-valued vector spaces. Some of the mathematical background may have more general definitions in terms of arbitrary fields, but we will present them just in terms of $\mathbb{R}$. We assume knowledge of measure-theoretic probability theory.

1.1.1 Markov Chains

The theory of Markov chains will be useful when considering reinforcement learning problems with finite state space. We describe Markov chains that take value on a set $E = \{1, \ldots, n\}$.

Definition 1.1 (Stochastic Matrix)**.** [8]. $T = \left(T_{x,y}\right)_{x,y \in E}$ is a stochastic matrix if each row of T is a probability density on E.

Proposition 1.2. *[4]. The eigenvalues of a stochastic matrix T are bounded in absolute value by 1.*

Definition 1.3 (Markov Chain)**.** [8]. A sequence $X_0, X_1, \ldots$ of random variables defined on a probability space $(\Omega, \Sigma, \mathbb{P})$ and taking values in E is called a *Markov chain* with state space E and transition matrix T if

$$\mathbb{P}\left(X_{n+1} = x_{n+1} | X_0 = x_0, \ldots, X_n = x_n\right) = T_{x_n, x_{n+1}}$$

for every $n \geq 0$ and $x_1, \ldots, x_{n+1} \in E$ with $\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n) > 0$.

Definition 1.4. [8]. A probability distribution μ such that $\mu \cdot T = \mu$ is called a stationary distribution.

Theorem 1.5 (Ergodic theorem for Markov Chains)**.** *[8]. Suppose there exists some $k \geq 1$ such that $T^k_{x,y} \geq 0$ for all $x, y \in E$. Then for every $y \in E$, the limit*

$$\lim_{n\to\infty} T^n_{x,y} = \mu(y) > 0$$

exists and does not depend on x. Furthermore, the limit μ is the unique probability density on E satisfying

$$\mu \cdot T = \mu \ .$$

Thus, μ is a stationary distribution.

1.1.2 Conditional Expectation and Conditional Probabilities

In this section we define some definitions that will be useful since we often consider conditional probabilities. The definitions in this section are drawn directly from [14].

Theorem 1.6. *Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space and $\Sigma_0 \subset \Sigma$ a σ-algebra. There exists a unique linear continuous map, $T : L^2(\Omega, \Sigma, \mathbb{P}) \to L^2(\Omega, \Sigma_0, \mathbb{P})$, such that for any $X \in L^2(\Omega, \Sigma, \mathbb{P})$ and $A \in \Sigma_0$,*

$$\int_A X d\mathbb{P} = \int_A T(X) d\mathbb{P} \ . \tag{1.1}$$

Definition 1.7 (Conditional Expectation)**.** The unique map T satisfying the properties of theorem 1.6 is called the conditional expectation with respect to the σ-algebra Σ_0. For $X \in L^2(\Omega, \Sigma, \mathbb{P})$, the conditional expectation of X given Σ_0 will be denoted by $\mathbb{E}\left(X|\Sigma_0\right) \coloneqq T(X)$.

Definition 1.8 (Conditioning Over a Random Variable)**.** Let $X_0 : \Omega \to \mathbb{R}$ be a random variable and $\mathcal{B}$ denote the Borel σ-algebra of $\mathbb{R}$. For $X \in L^2(\Omega, \Sigma, \mathbb{P})$, define

$$\mathbb{E}\left(X|X_0\right) \coloneqq \mathbb{E}\left(X|\sigma(X_0)\right) \ ,$$

where $\sigma(X_0) \coloneqq \left\{X_0^{-1}(B) : B \in \mathcal{B}\right\}$ is the σ-algebra generated by X_0. $\mathbb{E}(\cdot|X_0)$ is referred to as the conditional expectation given the random variable X_0.

Remark 1.9 (Conditioning Over an Event)**.** Consider the σ-algebra generated by an event $A_0 \in \Sigma$. This is given by $\sigma(A_0) := \{\emptyset, A_0, A_0^c, \Omega\}$. Define for $X \in L^2(\Omega, \Sigma, \mathbb{P})$

$$\mathbb{E}\left(X|A_0\right) := \frac{1}{\mathbb{P}\left(A_0\right)} \int_{A_0} X d\mathbb{P} \ . \tag{1.2}$$

This is called the conditional expectation of X given the event A_0 and represents the average value of X given that we know the event A_0 has occurred. We also define, for $B \in \Sigma$, the conditional probability of B given A_0,

$$\mathbb{P}\left(B|A_0\right) := \mathbb{E}(\mathbf{1}_B|A_0) = \frac{\mathbb{P}\left(B \cap A_0\right)}{\mathbb{P}\left(A_0\right)} \ . \tag{1.3}$$

$\mathbb{P}\left(B|A_0\right)$ is the probability of the event B occurring given that the event A_0 has occurred.

Proposition 1.10 (Conditional Expectation Properties)**.** *Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space and $\Sigma_0 \subset \Sigma$ a sub-σ-algebra. The following properties hold for any X and Y in $L^2(\Omega, \Sigma, \mathbb{P})$.*

(i) [Linearity]. For any $a, b \in \mathbb{R}$, $\mathbb{E}\left(aX + bY|\Sigma_0\right) = a\mathbb{E}\left(X|\Sigma_0\right) + b\mathbb{E}\left(Y|\Sigma_0\right)$.

(ii) [Monotonicity]. Suppose that $X \leq Y$ almost surely. Then, $\mathbb{E}\left(X|\Sigma_0\right) \leq \mathbb{E}\left(Y|\Sigma_0\right)$ almost surely.

(iii) [The Tower Property]. Let Σ_1 be a sub-σ-algebra of Σ_0. Then,

$$\mathbb{E}\left(\mathbb{E}\left(X|\Sigma_0\right)|\Sigma_1\right) = \mathbb{E}\left(X|\Sigma_1\right) = \mathbb{E}\left(\mathbb{E}\left(X|\Sigma_1\right)|\Sigma_0\right) \ .$$

(iv) [Jensen's Inequality]. For any convex $\phi : \mathbb{R} \to \mathbb{R}$,

$$\phi\left(\mathbb{E}\left(X|\Sigma_0\right)\right) \leq \mathbb{E}\left(\phi(X)|\Sigma_0\right) \ .$$

(v) [Law of Total Expectation].

$$\mathbb{E}\left(\mathbb{E}\left(X|Y\right)\right) = \mathbb{E}\left(X\right) \ .$$

1.1.3 Markov Processes on Arbitrary State Spaces

When considering the case of a continuous state space, we will often need more general theories about Markov processes. This section formalises the definitions and results in this case.

Definition 1.11. [14]. A discrete time filtration on the probability space $(\Omega, \Sigma, \mathbb{P})$ is a sequence of σ-algebras on Ω, $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ with the property that

$$\mathcal{F}_0 \subset \mathcal{F}_1 \subset \ldots \subset \Sigma \ .$$

Definition 1.12 (Stochastic processes)**.** [7]. Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space. If $\left(\mathcal{S}, \mathcal{B}\left(\mathcal{S}\right)\right)$ is a measurable space then a *$\mathcal{S}$-valued stochastic process* is a sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ such that

$$X_n : \Omega \to \mathcal{S} \ .$$

Definition 1.13 (Stationary Stochastic Process)**.** [7]. Let $(\mathcal{S}, \mathcal{B}(\mathcal{S}))$ be a measurable space denoted as the state space and $(\Omega, \Sigma, \mathbb{P})$ be a probability space. An $\mathcal{S}$-valued stochastic process $\{X_n\}_{n\in\mathbb{N}}$ is called *stationary* if $(X_1, X_2, \ldots)$ is equally distributed with $(X_k, X_{k+1}, \ldots)$ for any $k \in \mathbb{N}$.

Definition 1.14 (Filtration-adapted Stochastic Processes)**.** [7]. Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space with filtration $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$. A stochastic process $\{X_n\}_{n\in\mathbb{N}}$ is said to be *adapted* to the filtration $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ if for each $n \in \mathbb{N}$, X_{n+1} is $\mathcal{F}_n$-measurable.

Definition 1.15 ((Time-homogeneous) Markov Processes)**.** [7]. Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space with filtration $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ and let $\left(\mathcal{S}, \mathcal{B}(\mathcal{S})\right)$ be a measurable space. An $\mathcal{S}$-valued stochastic process adapted to the filtration $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ is called a Markov process if for any $B \in \mathcal{B}(\mathcal{S})$

$$\mathbb{P}\left(X_{n+1} \in B | \mathcal{F}_n\right) = \mathbb{P}\left(X_{n+1} \in B | X_n\right) \ . \tag{1.4}$$

We define the *transition probability kernel* of the process as $p(B|X_n) \coloneqq \mathbb{P}\left(X_{n+1} \in B | X_n\right)$.

If equation 1.4 is independent of n, i.e.

$$\mathbb{P}\left(X_{n+k+1} \in B | X_{n+k}\right) = \mathbb{P}\left(X_{k+1} \in B | X_k\right) \ , \quad \forall k \in \mathbb{N} \ ,$$

Then we say that the Markov process is *time-homogeneous*.

Definition 1.16 (Measure-Preserving Transformation)**.** [7]. Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space. A measurable map $T : \Omega \to \Omega$ is called a measure-preserving transformation if $\mathbb{P}(A) = \mathbb{P}(T^{-1}A)$ for all $A \in \Sigma$.

1.1.4 Ergodicity in Stationary Stochastic Processes

In this section we introduce a general set-up to formulate ergodic theory of stationary stochastic processes. We first begin by defining an ergodic probability measure. Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote a probability space and let $\Theta : \Omega \to \Omega$ denote a measure-preserving transformation on $(\Omega, \mathcal{F}, \mathbb{P})$, i.e.:

$$\mathbb{P}(A) = \mathbb{P}(\Theta^{-1}(A)) \quad \text{for all } A \in \mathcal{F}$$

Now let $\mathcal{J}$ denote the sub-σ-algebra of $\mathcal{F}$ containing all Θ-invariant events:

$$\mathcal{J} \coloneqq \{A \in \mathcal{F} : \Theta^{-1}(A) = A\}.$$

We can now define an ergodic probability measure as follows:

Definition 1.17 (Ergodic probability measure)**.** [7].
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and Θ be a measure-preserving transformation. Then $\mathbb{P}$ is called *ergodic* (with respect to Θ) if and only if any invariant event $A \in \mathcal{J}$ has probability zero or one.

Theorem 1.18 (Birkhoff's Ergodic Theorem)**.**
Suppose that Θ is a measure-preserving transformation and let $p \in [1, \infty)$. Then as $n \to \infty$,

$$\frac{1}{n}\sum_{i=0}^{n-1} F \circ \Theta^i \to \mathbb{E}[F|\mathcal{J}] \quad P\text{-almost surely in } L^p(\Omega, \mathcal{F}, \mathbb{P})$$

for any random variable $F \in L^p(\Omega, \mathcal{F}, \mathbb{P})$. In particular, if $\mathbb{P}$ is ergodic then

$$\frac{1}{n}\sum_{i=0}^{n-1} F \circ \Theta^i \to \mathbb{E}[F] \quad P\text{-almost surely in } L^p(\Omega, \mathcal{F}, \mathbb{P}).$$

[7]

Ergodicity in Stationary Markov Processes

Now consider the special case where X_n is a time-homogeneous Markov process on $(\Omega, \mathcal{F}, \mathbb{P})$ with transition kernel p. Let μ be a probability measure on $(\mathcal{S}, \mathcal{B}(\mathcal{S}))$. If μ satisfies

$$\mu(B) = \int p(y, B)\mu(y)dy \quad \text{for all } B \in \mathcal{B}(\mathcal{S})$$

then μ is called a stationary distribution. Define by P_μ the probability measure on $(\Omega, \mathcal{F})$ given by

$$P_\mu = \mu \cdot \prod_{n=1}^{\infty} p$$

P_μ represents the law of the Markov process X_n starting from an initial distribution μ. If μ is a stationary distribution for X_n, then X_n is a stationary process and the left-shift transformation Θ is measure-preserving w.r.t. P_μ. As noted in [7], in the case where P_μ is ergodic w.r.t Θ, Birkhoff's theorem straightforwardly implies the following two results.

Remark 1.19 (Estimating the transition kernel p)**.**
For any Borel sets $A, B \in \mathcal{B}(\mathcal{S})$,

$$\frac{1}{n}\sum_{t=0}^{n-1} \mathbf{1}_{A\times B}(X_t, X_{t+1}) \xrightarrow{n\to\infty} \mathbb{E}[\mathbf{1}_{A\times B}(X_0, X_1)] = \int_A \mu(s)p(s,B)ds \quad P_\mu\text{- a.s.}$$

Remark 1.20 (Law of large numbers)**.**
For any function $f \in L^1(\mathcal{S}, \mu)$,

$$\frac{1}{n}\sum_{t=0}^{n-1} f(X_t) \xrightarrow{n\to\infty} \int f d\mu \quad P_\mu - \text{a.s.}$$

Thus, Birkhoff's theorem implies that many quantities can be estimated effectively. In the case where a time-homogeneous Markov process begins in its stationary distribution and its law P_μ is ergodic w.r.t. the left-shift transformation, the transition kernel can be estimated as well as all random variables with finite expectation since $f \in L^1(\mathcal{S}, \mu) \implies \mathbb{E}(|f|) < \infty$.

1.1.5 Hilbert Space

We present some standard definitions and properties of Hilbert spaces in this section. These results are all drawn from [6] and a more detailed presentation can be found there.

Definition 1.21 (Hilbert Space)**.** A Hilbert space is a vector space $\mathcal{H}$ over $\mathbb{R}$ together with an inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ such that relative to the metric $\|\cdot\|_{\mathcal{H}} = \sqrt{\langle\cdot,\cdot\rangle_{\mathcal{H}}}$, $\mathcal{H}$ is a complete metric space. Furthermore, a Hilbert space is separable, that is there exists a countable collection $\{f_k\}$ of elements in $\mathcal{H}$ such that their linear combinations are dense in $\mathcal{H}$.

Remark 1.22. The space $L^2(\mu)$ with inner product $\langle f, g\rangle_\mu = \int fgd\mu$ and induced norm $\|f\|_\mu = \sqrt{\langle f, f\rangle_\mu}$ is a Hilbert space [6]. We will often look at the case of $L^2(\mathcal{S}, \mu)$, where $\mathcal{S}$ is a compact subset of $\mathbb{R}$. In this case, it is also a standard result that $L^2(\mathcal{S}, \mu)$ is also a Hilbert space.

Definition 1.23 (Orthogonality)**.** If $\mathcal{H}$ is a Hilbert space and $f, g \in \mathcal{H}$, then f and g are orthogonal if $\langle f, g\rangle = 0$. Often, we will denote this as $f \perp g$. If $A, B \subseteq \mathcal{H}$, then $A \perp B$ if $f \perp g$ for every f in A and g in B.

Remark 1.24. Having the concept of orthogonality is one of the greatest advantages of Hilbert spaces. In particular, we immediately get the familiar Pythagorean theorem.

Theorem 1.25 (Pythagorean Theorem)**.** *If $f_1, f_2, \ldots, f_n$ are pairwise orthogonal vectors in $\mathcal{H}$, then*

$$\|f_1 + f_2 + \ldots + f_n\|^2 = \|f_1\|^2 + \|f_2\|^2 + \ldots + \|f_n\|^2 \ .$$

We now detail the Riesz representation theorem. Its usage becomes important when dealing with adjoint operators on Hilbert space.

Definition 1.26 ((Bounded) Linear Functional)**.** A linear functional is a linear transformation from a hilbert space $\mathcal{H}$ to $\mathbb{R}$, i.e. $l : \mathcal{H} \to \mathbb{R}$. We say that l is a bounded linear functional if there exists a constant $c > 0$ such that $\left|l(f)\right| \leq c\|f\|$.

Theorem 1.27 (Riesz Representation Theorem)**.** *If $L : \mathcal{H} \to \mathbb{R}$ is a bounded linear functional then there is a unique $g \in \mathcal{H}$ such that*

$$l(f) = \langle f, g\rangle \quad \textit{and} \quad \|f\|_{\mathcal{H}} = \|g\|_{\mathcal{H}} \ , \qquad \textit{for all } f \in \mathcal{H}$$

1.1.6 Operators on Hilbert Space

Of much more interest to us are operators on Hilbert space. We begin by reproducing some standard definitions before moving toward properties of adjoint operators and oblique projection operators that are important in the proceeding discussions. Again, the results here can be found in [6].

Definition 1.28 (Bounded Linear Operator)**.** Let $\mathcal{H}_1, \mathcal{H}_2$ be Hilbert spaces and let $A : \mathcal{H}_1 \to \mathcal{H}_2$ be a linear operator. We say that A is a bounded linear operator if there exists a constant $c > 0$ such that $\|Ah\| \leq c\|h\|$ for all $h \in \mathcal{H}_1$.

Proposition 1.29. *Let $\mathcal{H}_1, \mathcal{H}_2$ be Hilbert spaces and let $A : \mathcal{H}_1 \to \mathcal{H}_2$ be a linear operator. The following are then equivalent:*

(a) A is continuous.

(b) There exists a constant $c > 0$ such that $\|Ah\| \leq c\|h\|$ for all $h \in \mathcal{H}_1$.

Thus from the proposition above it is clear that continuous linear operators and bounded linear operators are equivalent. In this section, let $B(\mathcal{H}_1, \mathcal{H}_2)$ denote the set of bounded linear operators of the form $A : \mathcal{H}_1 \to \mathcal{H}_2$.

Definition 1.30 (Adjoint Operator)**.** Let $\mathcal{H}$ be a Hilbert space and $T : \mathcal{H} \to \mathcal{H}$ be a bounded linear operator. Then there exists a unique bounded linear operator $T^* : \mathcal{H} \to \mathcal{H}$ on $\mathcal{H}$ such that:

(a) $\langle Tf, g\rangle = \langle f, T^*g\rangle$,

(b) $\|T\| = \|T^*\|$,

(c) $(T^*)^* = T$.

The bounded linear operator satisfying the above conditions, T^*, is called the adjoint of T.

As an important side note, proving the existence of such an adjoint operator proceeds by using the Riesz representation theorem [6]. Instinctively, the adjoint operator is a generalisation of the transpose of a matrix (a linear transformation in finite dimensions). An important property in our context is the relationship between the image and kernels of a bounded linear operator and its adjoint. This next result can be thought of as generalising the fundamental theorem of linear algebra.

Proposition 1.31. *Suppose $\mathcal{H}$ is a separable Hilbert space and let $P : \mathcal{H} \to \mathcal{H}$ be a bounded linear operator and P^* its adjoint. Then*

$$ker(P) \perp im(P^*)$$
$$im(P) \perp ker(P^*) \ .$$

Proof. To see the first condition, suppose that $x \in ker(P)$ and $y \in im(P^*)$. Since $y = P^*z$ for some $z \in \mathcal{H}$, we have

$$\begin{aligned} \langle x, y\rangle &= \langle x, P^*z\rangle \\ &\overset{(a)}{=} \langle Px, z\rangle \\ &\overset{(b)}{=} 0 \ , \end{aligned}$$

where (a) follows by the definition of the adjoint and (b) follows since $x \in ker(P)$. Thus $ker(P) \perp im(P^*)$. The second condition follows a similar line of argument. Let $x \in ker(P^*)$ and $y \in im(P)$. Then for some $z \in \mathcal{H}$, $y = Pz$. Thus

$$\begin{aligned} \langle x, y \rangle &= \langle x, Pz \rangle \\ &\overset{(a)}{=} \langle Pz, x \rangle \\ &\overset{(b)}{=} \langle z, P^* x \rangle \\ &\overset{(c)}{=} 0 \end{aligned}$$

where (a) follows as $\langle \cdot, \cdot \rangle$ is symmetric, (b) follows by definition of the adjoint, and (c) follow since $x \in ker(P^*)$. □

We now provide a definition of oblique projection operators. The properties of these operators will be of principal importance in the remainder of this thesis.

Definition 1.32 (Oblique Projection Operators)**.** Let $\mathcal{H}$ be a Hilbert space and $\Pi : \mathcal{H} \to \mathcal{H}$ a bounded linear operator.Π is an oblique projection operator if Π is idempotent, i.e. $\Pi^2 = \Pi$, and Π admits a direct sum decomposition of $\mathcal{H}$

$$\mathcal{H} = im(\Pi) \oplus ker(\Pi) \ .$$

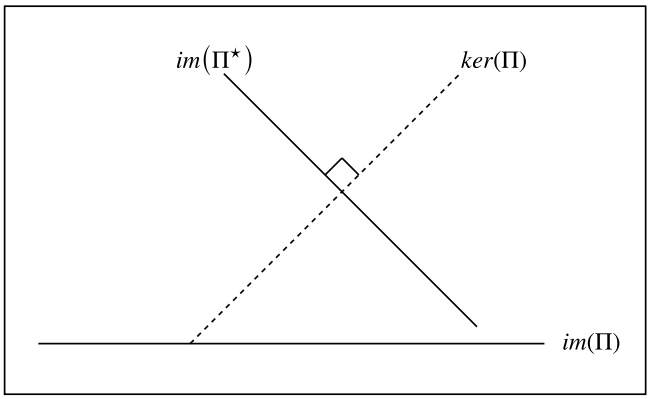

Figure 1.1: A simple $\mathbb{R}^2$ example of the geometry of an oblique projection.

To gain some intuition towards oblique projections, consider an oblique projection Π on $\mathbb{R}^2$ visualised in Figure 1.1. Clearly in this case, even though the kernel and image of Π are not orthogonal, they produce a direct sum decomposition of $\mathbb{R}^2$. Π can be thought of as projecting any vector in $\mathbb{R}^2$ *along* its kernel onto its

image. Equivalently, Π can be thought of as projecting orthogonally to $im(\Pi^*)$. Then for any $v \in \mathbb{R}^2$, $v - \Pi v \in ker(\Pi)$ and thus $v - \Pi v \perp im(\Pi^*)$.

1.1.7 Additional Useful Definitions and Results

The definitions presented here will be used sparingly throughout the thesis.

Definition 1.33. [12]. A *convex set* Ω is a set where for any $x_0, x_1 \in \Omega$, the line segment joining them also lies in Ω. Formally,

$$tx_1 + (1-t)x_0 \in \Omega \ , \quad \forall 0 \leq t \leq 1 \ .$$

Definition 1.34 (Coercive Function)**.** [12]. A function $f : \mathbb{R}^n \to \mathbb{R}$ is called *coercive* if $f(x) \to +\infty$ as $|x| \to \infty$.

Definition 1.35 (Convex Function)**.** [12]. A function $f : \Omega \to \mathbb{R} \cup \{-\infty, +\infty\}$ where Ω is a convex set in $\mathbb{R}^n$ is *convex* if

$$f(tx_1 + (1-t)x_0) \leq tf(x_1) + (1-t)f(x_0)$$

for all $x_1, x_0 \in \Omega$ and $0 \leq t \leq 1$.

Chapter 2

Introduction to Reinforcement Learning

As the title suggest we look to introduce readers to the mathematical foundations of reinforcement learning in this chapter. We begin by specifying the agent-environment framework before specialising into a concrete example of an environment known as a Markov decision process. The theory of dynamic programming is then covered as it forms the foundation for solving most reinforcement learning problems.

2.1 The Agent-Environment Framework

Reinforcement learning problems are typically framed as a dynamic decision problem between an agent acting out decisions and an environment returning a re-

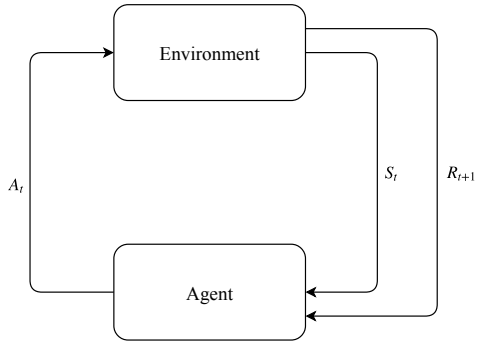

Figure 2.1: An abstract view of the agent-environment interaction.

sponse. The *agent* can take actions $a \in \mathcal{A}$. For instance, if the agent is a self-driving car, the actions could correspond to accelerate, brake, and the degree of the steering wheel. In the cases we consider, the *environment* returns a state $s \in \mathcal{S}$ and a reward $r \in \mathbb{R}$. Considering the self-driving car example again, the state could correspond to the position of the car and the reward as the negative of the distance from the destination. In the next section, where we consider a particular type of agent-environment interaction, we will define the sets $\mathcal{S}$ and $\mathcal{A}$ explicitly. The agent and the environment interact in cycles. At time t, an agent receives a state s_t from the environment and decides upon an action a_t to take. After taking action a_t, the environment responds and provides the agent with a reward r_t and a new state s_{t+1}. The agent then decides upon its next action. In this manner, the agent and environment interact in a cycle to produce a sequence indexed at discrete time steps:

$$s_0, a_0, r_0, s_1, a_1, r_1, s_2, a_2, r_2, \ldots$$

In what follows, it will be conventional to consider infinitely long interactions between the agent and the environment. The set of all such sequences is $(\mathcal{S} \times \mathcal{A} \times \mathbb{R})^{\mathbb{N}}$. A sequence in $(\mathcal{S} \times \mathcal{A} \times \mathbb{R})^{\mathbb{N}}$ can be viewed as the realisation of a sequence of random variables:

$$S_0, A_0, R_0, S_1, A_1, R_1, S_2, A_2, R_2, \ldots$$

where at time t, S_t is an $\mathcal{S}$-valued random variable, A_t is an $\mathcal{A}$-valued random variable, and R_t is an $\mathbb{R}$-valued random variable. We denote this sequence of random variables as the *interaction sequence*. In particular, for a sequence $\omega = (s_t, a_t, r_t)_{t\in\mathbb{N}}$ in $(\mathcal{S} \times \mathcal{A} \times \mathbb{R})^{\mathbb{N}}$, we have at time t

$$S_t(\omega) = s_t \ , \quad A_t(\omega) = a_t \ , \quad R_t(\omega) = r_t \ .$$

Typically, the initial state S_0 is provided by a distribution that may be independent of the underlying environment process. We will denote the sub-sequences of the interaction sequence $\{S_t\}_{t\in\mathbb{N}}$, $\{A_t\}_{t\in\mathbb{N}}$, and $\{R_t\}_{t\in\mathbb{N}}$ as the *state sequence, action sequence*, and *reward sequence* respectively. In the situations we consider, the state and action sequences evolve stochastically, whilst the reward received at time t, viewed as the *expected reward*, will be a function of the state S_t and A_t, that is, for some function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$

$$S_0, A_0, R(S_0, A_0), S_1, A_1, R(S_1, A_1), S_2, A_2, R(S_2, A_2), \ldots$$

We note that generalising to the case of a stochastic reward sequence is routine and interested readers, please see [20]. Given an interaction sequence up to time t, the goal in reinforcement learning is to find a probability distribution, known as a *policy*, $\pi : (\mathcal{S} \times \mathcal{A} \times \mathbb{R})^t \times \mathcal{A} \to [0, 1]$ that determines how an agent will choose its next action. We note that the agent-environment interaction can be represented mathematically in a very general way that covers many practical scenarios. However, we will restrict our attention to a particular type of interaction that satisfies the Markov property known as Markov decision processes.

2.2 Markov Decision Processes

An MDP is given by a tuple $M = (\mathcal{S}, \mathcal{A}, R, T)$ where each variable represents the *state space*, *action space*, the *expected reward function* and the *transition function* respectively. We take the *state space* $\mathcal{S}$ to be a compact subspace of $\mathbb{R}$ with the Borel σ-algebra $\mathcal{B}(\mathcal{S})$ such that $(\mathcal{S}, \mathcal{B}(\mathcal{S}))$ is a measurable space. Let the *action space* $\mathcal{A}$ be a finite set that, coupled with its power-set, forms a measurable space $(\mathcal{A}, \mathcal{P}(\mathcal{A}))$. The *expected reward function* is a function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and, as the name suggests, represents the expected reward to be received for a given state and action. We now look to derive the transition function and policy generating the action and state sequences.

Consider the subsequence $(S_t, A_t)_{t\in\mathbb{N}}$, denoted as the *state-action sequence*, in the agent-environment interaction sequence. For brevity, let $\mathcal{X}$ be defined as $\mathcal{X} \coloneqq \mathcal{S} \times \mathcal{A}$ and its corresponding σ-algebra as $\mathcal{B}(\mathcal{X}) \coloneqq \mathcal{B}(\mathcal{S}) \times \mathcal{P}(\mathcal{A})$. The state-action sequence can be modelled as a discrete, time-homogeneous Markov process $\{X_t\}_{t\in\mathbb{N}}$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where $\Omega = \mathcal{X}^{\mathbb{N}}$. For any $t \in \mathbb{N}$ and $\omega \in \Omega$,

$$X_t(\omega) = (S_t, A_t)$$

where S_t, A_t are random variables taking values in $\mathcal{S}$ and $\mathcal{A}$ respectively. The filtration generated by $\{X_t\}_{t\in\mathbb{N}}$ at time t is given by

$$\mathcal{F}_t \coloneqq \sigma\big(X_s : s \leq t\big) \ .$$

It is clear to see that $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \ldots \subset \mathcal{F}$ and that at time t, X_t is $\mathcal{F}_t$-measurable. As a Markov process, $\{X_t\}_{t\in\mathbb{N}}$ satisfies the *Markov property*:

$$\mathbb{P}\big(\{X_{t+1} \in B\}\big|\mathcal{F}_t\big) = \mathbb{P}\big(\{X_{t+1} \in B\}\big|X_t\big) \ , \forall t \in \mathbb{N}, \ B \in \mathcal{B}(\mathcal{X}) \ .$$

Furthermore, since $\{X_t\}_{t\in\mathbb{N}}$ is time-homogeneous, the transition probabilities are independent of the time-step, that is

$$\mathbb{P}\big(\{X_{t+k+1} \in B\}\big|X_{t+k}\big) = \mathbb{P}\big(\{X_{k+1} \in B\}\big|X_k\big) \ , \forall t, k \in \mathbb{N} \ .$$

The *transition kernel* of $\{X_t\}_{t\in\mathbb{N}}$ is defined as $p(B|X_k) \coloneqq \mathbb{P}\big(\{X_{k+1} \in B\}\big|X_k\big)$. We note for an initial distribution ν over X_0, p and ν completely determine the joint distribution of $X_0, ..., X_k$ for any k. Indeed, for all $k \in \mathbb{N}$ we have by the law of total probability and the Markov property

$$\mathbb{P}(X_0 \in B_0, \ldots, X_k \in B_k) = \int_{x_0\in B_1} \ldots \int_{x_k\in B_k} \nu(dx_0)p(dx_1|x_0)\ldots p(dx_k, x_{k-1}) \ .$$

To justify the existence of such a Markov process, it suffices to apply Kolmogorov's extension theorem with the set of all finite-dimensional joint distributions defined above [4].

Let λ denote the product measure comprised of the Lebesgue measure over $\mathcal{S}$ and the counting measure over $\mathcal{A}$. We assume that the transition kernel emits a density function P , that is p is given by

$$p(B|x) = \int_B P(y|x)\lambda(dy) \quad \forall x \in \mathcal{X}, B \in \mathcal{B}(\mathcal{X})$$

where $P : \mathcal{X} \times \mathcal{X} \to [0, 1]$ is a measurable function satisfying

$$\int_{\mathcal{X}} P(y|x)\lambda(dy) = 1 \ .$$

We refer to P as the *transition density function* with respect to the measure λ. For brevity, we will henceforth drop λ from the integrals noting that unless a different measure is explicitly specified, either the Lebesgue or counting measure is taken depending on the cardinality of the underlying set. The transition density function P can be considered from a different perspective by expanding the arguments. Let $x = (s, a), y = (s', a') \in \mathcal{X}$. Then

$$P(y|x) = P(S_{k+1} = s', A_{k+1} = a'|S_k = s, A_k = a) \ .$$

Moving forward, we will denote $P(y|x)$ by $P(s', a'|s, a)$. By taking the chain rule on conditional probabilities, we have

$$P(s', a'|s, a) = P(a'|s', s, a)P(s'|s, a) \ .$$

We denote the *transition function* as $T(s'|s, a) \coloneqq P(s'|s, a)$. A policy specified by the MDP is given by $\pi(a'|s', s, a) \coloneqq P(a'|s', s, a)$. In what follows, we assume

that the policies are Markov, that is the next action a' chosen by π is independent of s, a given s'. Thus $\pi(a'|s') = \pi(a'|s', s, a)$. An immediate consequence is that for any choice of transition function and policy π we can recover the transition density kernel of the underlying state-action sequence. As a result, an MDP and a policy fully characterise an interaction sequence.

In what follows, we assume that the expected reward function is $\|\cdot\|_\infty$-bounded by some finite quantity $\mathcal{G} > 0$.

Assumption 1. Assume that the expected reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is $\|\cdot\|_\infty$-bounded by a finite value $\mathcal{G} > 0$. That is

$$\|R\|_\infty < \mathcal{G} < \infty \ .$$

Remark 2.1. It can be seen immediately that the reward sequence is bounded by $\mathcal{G}$ at each time step, that is $|R_t| < \mathcal{G}$ for all $t \in \mathbb{N}$.

The transition function T and the expected reward function R determine the way in which the MDP environment can respond to the agent's actions and the previous state. Thus given a random initial state S_0, the transition function, expected reward function and policy determine the Markov process $\{X_n\}_{n\in\mathbb{N}}$ and hence, the interaction sequence.

2.2.1 Stationary Policies

A *stationary policy* π defines a distribution over $\mathcal{A}$ for a given state in $\mathcal{S}$ that is time invariant. We denote by $\pi(\cdot|s)$ the probability density over $\mathcal{A}$ in state s. If a stationary policy is followed by an agent interacting with an MDP, then the action sequence is distributed according to π, that is for all $t \in \mathbb{N}$

$$A_t \sim \pi(\cdot|S_t) \ .$$

A *deterministic stationary policy* is a stationary policy π mapping states directly to actions, $\pi : \mathcal{S} \to \mathcal{A}$). In the case where $\mathcal{A}$ is finite, we can think of deterministic stationary policies as being a stationary policy with point mass. In what follows, it will be convenient to work with deterministic stationary policies and so we define by convention

$$\pi(S_t) = A_t \qquad \text{if } \pi(A_t|S_t) = 1 \ .$$

For an interaction sequence generated between a stationary policy and an MDP, the state sequence $\{S_t\}_{t\in\mathbb{N}}$ forms a time-homogeneous Markov process. We denote by Π_{stat} the set of stationary policies. In what follows, we will only consider

stationary deterministic policies defined under our convention as they are more convenient in developing the underlying theory of dynamic programming. Generalisation to stochastic policies is again routine [20]. Hence we will refer to deterministic stationary policies as policies for brevity.

2.2.2 MDPs for fixed policy

In future sections, the evolution of the state sequence generated by an agent following a fixed policy and an MDP is often of interest. When the policy is fixed, the state sequence can be modelled as a Markov process. Suppose we are given a policy π and an MDP $M = (\mathcal{S}, \mathcal{A}, \mathbb{R}, \mathcal{T})$. Then we can define the transition function depending on π as the following:

$$T^{\pi}(s'|s) = T(s'|s, \pi(s)) \ .$$

The expected reward function depending on π can also be expressed similarly as

$$R^{\pi}(s) = R(s, \pi(s)) \ .$$

Combined with an initial state S_0, the state sequence can be viewed as a time-homogeneous Markov process with transition kernel defined by

$$P^{\pi}(B|x) = \int_B T^{\pi}(y|x)dy \ , \quad \forall B \in \mathcal{B}(\mathcal{S}), \ x \in \mathcal{S} \ .$$

We will often consider the dynamics of the underlying reinforcement problem in the case of a fixed policy and so the functions T^{π} and R^{π} will be used extensively.

2.2.3 Value Functions

The value function of a policy summarises the expected discounted reward to be received from a given initial state for a fixed policy. The discount rate is given by $\gamma \in (0, 1)$ and weights rewards in later time steps of the reward sequence less than rewards received more immediately. As we will see, considering value functions also provides a mechanism for finding optimal policies.

Consider a fixed deterministic policy π that generates an interaction sequence with MDP M and initial distribution μ. We define the *value function* of π as follows.

Definition 2.2. For a discount rate $\gamma \in (0,1)$ and initial state s sampled from μ, the *value function* of π, $V^\pi : \mathcal{S} \to \mathbb{R}$ is given by

$$V^\pi(s) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t R_t \middle| S_0 = s\right] \tag{2.1}$$

where $\mathbb{E}_\pi$ denotes the expectation taken over T^π at each time step

Equivalently, equation (2.1) can be expressed as

$$V^\pi(s) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t R^\pi(S_t) \middle| S_0 = s\right] \tag{2.2}$$

and this is the representation that we will use more frequently. In order to ensure that the conditional expectation is well defined, we only consider initial distributions μ where $\mu(s) > 0$ for all $s \in \mathcal{S}$. The *action-value function*, often denoted the *Q-value* function, is a value function $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ associated with π. Assume that there is an initial distribution ν over the action space by which A_0 is selected randomly and $\nu(a) > 0$ for all $a \in \mathcal{A}$. Note that ν may be given by the policy π itself. Then the Q-value function for π is defined by

$$Q^\pi(s,a) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t R^\pi(S_t) \middle| S_0 = s, A_0 = a\right], \quad \forall s \in \mathcal{S}, a \in \mathcal{A}\ . \tag{2.3}$$

Methods that solve for the Q-value function, such as Q-learning [22], are quite popular in practice. However we restrict our attention to methods that solve for the value function as it is a related and equivalent task.

We can now consider the optimal value function over all stationary policies. Let $\leq$ define a partial order over the value functions, that is for any $V, \hat{V} \in \mathbb{R}^{\mathcal{S}}$,

$$V \leq \hat{V} \iff V(s) \leq \hat{V}(s)\ \forall s \in \mathcal{S}\ .$$

The optimal value function denotes the maximum possible expected return from a given initial state s. Let $V^* : \mathcal{S} \to \mathbb{R}$ denote the optimal value function over all stationary policies. Then V^* is well defined and is given by

$$V^*(s) = \sup_{\pi \in \Pi_{stat}} V^\pi(s)\ , \forall s \in \mathcal{S}\ . \tag{2.4}$$

We say that any policy that achieves the optimal value in *all* states, that is $V^\pi(s) = V^*(s)$ for all $s \in \mathcal{S}$, is *optimal*. Given our assumption that all reward values are bounded, we note that all value functions of stationary policies are bounded.

Proposition 2.3. *For any stationary policy π, its corresponding value function V^π is bounded in the infinity norm by $\frac{\mathcal{G}}{1-\gamma}$.*

Proof. For any $s \in \mathcal{S}$ and a fixed policy π we have

$$\begin{aligned} V^\pi(s) &= \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t R^\pi(S_t) \middle| S_0 = s\right] \\ &\overset{(a)}{\leq} \mathcal{G} \cdot \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t \middle| S_0 = s\right] \\ &= \frac{\mathcal{G}}{1-\gamma} < \infty \end{aligned}$$

where (a) follows by our assumption on R_t being bounded by $\mathcal{G}$. Thus, $\|V^\pi\|_\infty < \frac{\mathcal{G}}{1-\gamma}$ which is finite. □

We consider next the main framework for determining value functions.

2.3 Dynamic Programming

The mathematical foundations of dynamic programming provide the main theoretical framework for solving the reinforcement learning problem. In a broad sense, dynamic programming can be thought of as solving a sequential system of optimization problems. In what follows, we will present the theory for fixed deterministic policies formally whilst discussing the optimal case more informally. Readers interested in a more rigorous approach to the optimal case should turn to [20].

2.3.1 The Bellman Equations

The Bellman equations provide a different representation of V^π as a fixed point equation that is useful for solving for V^π. We first describe the Bellman equation for deterministic policies.

Lemma 2.4 (Bellman's Equations for Deterministic Policies)**.**
Let $\pi \in \Pi_{stat}$ be a deterministic policy and $s \in \mathcal{S}$ be any state. Then the value function V^π satisfies

$$V^\pi(s) = R^\pi(s) + \gamma \int_{\mathcal{S}} T^\pi(s'|s) V^\pi(s') ds' \ . \tag{2.5}$$

Proof. For a given starting state $s \in \mathcal{S}$, we can perform the following derivation on V^π:

$$\begin{aligned}
V^\pi(s) &= \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t R^\pi(S_t)\middle| S_0 = s\right] \\
&= \mathbb{E}_\pi\left[R^\pi(s) + \gamma\sum_{t=1}^{\infty}\gamma^{t-1}R^\pi(S_t)\middle| S_0 = s\right] \\
&\overset{(a)}{=} R^\pi(s) + \mathbb{E}_\pi\left[\gamma\sum_{t=1}^{\infty}\gamma^{t-1}R^\pi(S_t)\middle| S_0 = s\right] \\
&\overset{(b)}{=} R^\pi(s) + \mathbb{E}_{T^\pi}\left[\mathbb{E}_\pi\left[\gamma\sum_{t=1}^{\infty}\gamma^{t-1}R^\pi(S_t)\middle| S_0 = s, S_1\right]\middle| S_0 = s\right] \\
&\overset{(c)}{=} R^\pi(s) + \gamma\int_{\mathcal{S}} T^\pi(s'|s)\ \mathbb{E}_\pi\left[\sum_{t=1}^{\infty}\gamma^{t-1}R^\pi(S_t)\middle| S_0 = s, S_1 = s'\right]ds' \\
&\overset{(d)}{=} R^\pi(s) + \gamma\int_{\mathcal{S}} T^\pi(s'|s)\ \mathbb{E}_\pi\left[\sum_{t=1}^{\infty}\gamma^{t-1}R^\pi(S_t)\middle| S_1 = s'\right]ds' \\
&\overset{(e)}{=} R^\pi(s) + \gamma\int_{\mathcal{S}} T^\pi(s'|s)\ V^\pi(s')ds' \ .
\end{aligned}$$

We note that (a) follows since R^π is deterministic. In (b) we define E_{T^π} as the expectation taken over the transition function T^π in one step and the expression follows from the tower property (see preliminaries). The equality (d) follows since the expectation no longer has terms dependent on S_0. In (e), we used the definition of V^π to ultimately arrive at the expression in equation (2.5).

□

The relationship defined in equation (2.5) is known as the *Bellman equations*. We now define the accompanying Bellman operator. Note that $\mathbb{R}^{\mathcal{S}}$ denotes the set of all functions from $\mathcal{S}$ to $\mathbb{R}$.

Definition 2.5. For a deterministic policy π, the Bellman operator $\mathcal{T}^\pi : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$ is given by the following. For all $s \in \mathcal{S}$

$$(\mathcal{T}^\pi V)(s) = R^\pi(s) + \gamma\int_{\mathcal{S}} T^\pi(s'|s)V(s')ds' \ .$$

Thus the Bellman equation (2.5) can be written in compact form as the fixed point equation

$$V^\pi = \mathcal{T}^\pi V^\pi \ . \tag{2.6}$$

Note that $\mathcal{T}^\pi$ is an affine linear operator.

A key property of $\mathcal{T}^\pi$ is that it is a contraction with respect to $\|\cdot\|_\infty$.

Theorem 2.6. *The Bellman operator $\mathcal{T}^\pi$ is a $\|\cdot\|_\infty$-contraction.*

Proof. Let $\hat{V}, V \in \mathbb{R}^{\mathcal{S}}$ be two arbitrary functions mapping the state space to the real numbers. Then

$$\begin{aligned}
\left\|\mathcal{T}^\pi \hat{V} - \mathcal{T}^\pi V\right\|_\infty &= \gamma \sup_{s\in\mathcal{S}} \left| \int_{\mathcal{S}} T^\pi(s'|s) \left(\hat{V}(s') - V(s')\right) ds' \right| \\
&\leq \gamma \sup_{s\in\mathcal{S}} \int_{\mathcal{S}} T^\pi(s'|s) \left| \left(\hat{V}(s') - V(s')\right) \right| ds' \\
&\leq \gamma \sup_{s\in\mathcal{S}} \int_{\mathcal{S}} T^\pi(s'|s) \left\|\hat{V} - V\right\|_\infty ds' \\
&\overset{(a)}{=} \gamma \left\|\hat{V} - V\right\|_\infty
\end{aligned}$$

where (a) follows since $\int_{\mathcal{S}} T^\pi(s'|s)ds' = 1$. □

Thus, to find V^π, Banach's fixed point theorem guarantees us that starting from any initial function V_0, the sequence $\mathcal{T}^\pi V_0, (\mathcal{T}^\pi)^2 V_0, ...$ converges to the fixed point V^π. As we discuss in section 2.3.2, the value iteration and policy iteration algorithms exploit this property to find the solution.

We now provide a less rigorous discussion on how to find the optimal value functions. Reader's interested in the optimal case should consult [20]. The first key property is that the optimal value function can also be found as the fixed point of a fixed point equation.

Lemma 2.7 (Bellman's Optimal Equations)**.** *The optimal value function is the unique fixed point of the following fixed point equation*

$$V^*(s) = \max_{a\in\mathcal{A}} \left\{ R(s,a) + \gamma \int_{\mathcal{S}} T(s'|s,a) V^*(s') ds' \right\} . \tag{2.7}$$

Proof. See [20]. □

We similarly define the Bellman optimality operator.

Definition 2.8. The Bellman optimality operator $T^* : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$ is given by

$$(\mathcal{T}^* V)(s) = \max_{a\in\mathcal{A}} \left\{ R(s,a) + \gamma \int_{\mathcal{S}} T(s'|s,a) V(s') ds' \right\} .$$

The Bellman optimal equations can then be expressed succinctly as the following fixed point equation

$$\mathcal{T}^* V^* = V^* \ .$$

It can be shown that $\mathcal{T}^*$ is also a contraction with respect to $\|\cdot\|_\infty$ [20]. Thus, by Banach's fixed point theorem, the sequence $\mathcal{T}^* V_0, (\mathcal{T}^*)^2 V_0, ...$ converges to the optimal value function V^*.

Remark 2.9. As a consequence of the definition of T^*, for any $V \in \mathbb{R}^{\mathcal{S}}$ and $\pi \in \Pi_{stat}$, we have $T^* V \geq T^\pi V$.

Recall that the purpose of finding the value functions was to evaluate policies. The next theorem tells us how to find the optimal policy. We first define the notion of a *greedy policy*.

Definition 2.10. A policy π is greedy with respect to $V \in \mathbb{R}^{\mathcal{S}}$ if $\mathcal{T}^\pi V = \mathcal{T}^* V$.

Theorem 2.11. *Let V be the fixed point of $\mathcal{T}^*$ and assume that there is a policy π which is greedy with respect to V. Then $V = V^*$ and π is an optimal policy.*

Proof. See [20]. □

In the next section we describe the two main algorithms used to solve for the value function.

2.3.2 Value and Policy Iteration

Value iteration and policy iteration are the two main methods used to find the value function. Both techniques rely heavily upon the contraction property of $\mathcal{T}^\pi$ and $\mathcal{T}^*$ to generate the correct solution.

The value iteration method performs a fixed point iteration to generate a sequence of value functions that converges to the fixed point of (2.6). Given $\mathcal{T}^\pi$ and any initial value function V_0^π, value iteration generates the sequence $\{V_k^\pi\}_{k \geq 0}$ via

$$V_{k+1}^\pi \coloneqq \mathcal{T}^\pi V_k^\pi \ , \quad \forall k \geq 0 \ .$$

By Banach's fixed point theorem, the sequence is guaranteed to converge at a geometric rate. A similar sequence can be generated by value iteration using $\mathcal{T}^*$ to find the optimal value function.

The policy iteration algorithm generates a sequence of suboptimal policies that ultimately converges to the optimal policy. Suppose we start with a policy $\pi_0 \in \Pi_{stat}$ and generate the sequence of policies $\{\pi_k\}_{k\geq 0} \subseteq \Pi_{stat}$. For a given policy π_k, we perform a *policy evaluation step* to find V^{π_k} that satisfies

$$V^{\pi_k} = \mathcal{T}^{\pi_k} V^{\pi_k} \ .$$

Once V^{π_k} is found, we find π_{k+1} via a *policy improvement step* as follows:

$$\forall s \in \mathcal{S}, \quad \pi_{k+1}(s) = \arg\max_{a\in\mathcal{A}} \left[R(s,a) + \gamma \int_{\mathcal{S}} T(s'|s,a) V^{\pi_k}(s') ds'\right] \ . \tag{2.8}$$

If $\pi_{k+1} = \pi_k$, the algorithm terminates. The next theorem given in [20] establishes the convergence property of policy iteration.

Theorem 2.12. *Let π_0 be a policy with value function V^{π_0} and suppose that π is a policy that is greedy with respect to V^{π_0}, i.e. $T^{\pi} V^{\pi_0} = T^{*} V^{\pi_0}$. Then $V^{\pi} \geq V^{\pi_0}$. In particular, if $T^{*} V^{\pi_0}(s) > V^{\pi}(s)$ for some state s then π strictly improves upon π_0 at s: $V^{\pi}(s) > V^{\pi_0}(s)$.*

Proof. By definition, $T^{*} V^{\pi_0} \geq T^{\pi_0} V^{\pi_0}$. Then since π is greedy and V^{π_0} is the fixed point of $T^{\pi_0} V^{\pi_0}$, we have that $T^{\pi} V^{\pi_0} \geq V^{\pi_0}$. Applying T^{π} to both sides gives

$$\left(T^{\pi}\right)^2 V^{\pi_0} \geq T^{\pi} V^{\pi_0} \ .$$

Thus, $\left(T^{\pi}\right)^2 V^{\pi_0} \geq V^{\pi_0}$. Continuing in this fashion, we have for any $n \geq 0$ that

$$\left(T^{\pi}\right)^n V^{\pi_0} \geq V^{\pi_0} \ .$$

Then taking the limit as $n \to \infty$ on both sides gives $V^{\pi} \geq V^{\pi_0}$. Then since $T^{*} V^{\pi_0} \geq T^{\pi_0} V^{\pi_0}$, the second point follows as a direct consequence of the above. □

Both value iteration and policy iteration are quite slow in practice. Both methods require a precise model of the environment dynamics, i.e. the state transition functions and expected reward functions must be given. Many more practical algorithms do not operate with explicit state transition and expected reward functions, but instead are based on computing sample estimates. However, as we will see in later chapters, many of the most widely used algorithms look to approximate the behaviour of value iteration or policy iteration in some way.

2.4 The Case of a Finite State Space

So far in this chapter, most of the theory for reinforcement learning has been developed in the case where $\mathcal{S}$ is a compact subspace of $\mathbb{R}$. However, generalising these results to the countable case is fairly straightforward. For example, integrals over $\mathcal{S}$ are instead taken with respect to the counting measure and $\mathcal{B}(\mathcal{S})$ is taken to be the power-set. Density functions simplify to mass functions and the transition function T^π, expected reward function R^π and the value functions can be all be expressed using matrices and vectors. For simplicity, we will without loss of generality let $\mathcal{S} = \{1, \dots, n\}$ in the finite case. We present each of the mathematical objects under a finite state space and a fixed policy π. For brevity, the derivations are left out. Readers interested in the derivations are instead referred to [20].

Definition 2.13 (MDPs under a Finite State Space)**.** Suppose $\mathcal{S} = \{1, \dots, n\}$. Then we have the following definitions:

- The transition function is given by a matrix $T^\pi \in \mathbb{R}^{n \times n}$.
- The expected reward function is given by a vector $R^\pi \in \mathbb{R}^n$.
- The value function is given by a vector $V^\pi \in \mathbb{R}^n$.
- The Bellman operator is now a mapping from $\mathbb{R}^n$ to $\mathbb{R}^n$ and can be explicitly expressed as

$$\mathcal{T}^\pi V = R^\pi + \gamma T^\pi V \ , \quad \forall V \in \mathbb{R}^n \ .$$

Chapter 3

Approximate Solution Methods

Dynamic programming methods run into two main problems in practice: the transition function is often a-priori unknown and when the state space is 'large', possibly infinite, its methods become computationally intractable. In this chapter we focus on methods that look to solve the second issue. Typically, function approximation techniques are combined with pre-existing reinforcement learning algorithms to solve this issue. This chapter begins by defining some extra geometrical structure upon the value function space to work with before giving a brief description of linear function approximation. Projection equation methods will then be characterised in general. We then present a new perspective to characterising the solution of a projected Bellman equation. The chapter culminates in the definition of natural algorithms, which are the set of algorithms that we consider for our divergence results.

The Value Function Space Under Ergodicity

Recall that for a fixed policy π and MDP M we can view the state sequence as a Markov process with transition function T^π. Throughout this chapter we will assume that the state Markov process is ergodic and admits a stationary distribution. As is a natural assumption for practical purposes in reinforcement learning, we will also assume that the steady-state distribution of all states $s \in \mathcal{S}$ is greater than 0. These assumptions are summarised in the following.

Assumption 2. Assume that the state sequence $\{S_t\}_{t\in\mathbb{N}}$ is ergodic and admits a stationary distribution μ such that $\mu(s) > 0$ for all $s \in \mathcal{S}$.

Under this assumption, the space in which the value functions live inherits extra geometrical structure with respect to an inner product defined by μ. For

any $f, g \in \mathbb{R}^{\mathcal{S}}$,

$$\langle f, g \rangle_\mu = \int_{\mathcal{S}} f(s)g(s)\mu(s)ds \ .$$

To show that $\langle \cdot, \cdot \rangle_\mu$ is an inner product is routine. Furthermore, we define the induced norm $\|x\|_\mu = \sqrt{\langle x, x \rangle_\mu}$ as the μ-weighted quadratic norm. Then let $L^2(\mathcal{S}, \mu)$ define the set of functions in $\mathbb{R}^{\mathcal{S}}$ with finite μ-weighted quadratic norm:

$$L^2(\mathcal{S}, \mu) \coloneqq \{f \in \mathbb{R}^{\mathcal{S}} : \|f\|_\mu^2 < \infty\}.$$

We note that this is a separable Hilbert space (see preliminaries). For any stationary policy π, the value function V^π trivially resides within $L^2(\mathcal{S}, \mu)$ as $\|V^\pi\|_\infty < \infty$ (see Proposition 2.3). With respect to the extra structure provided by the inner product, we are now able to consider projections.

3.1 Linear Function Approximation

Linear function approximations are an important class of function approximation techniques due to their computational tractability. In most typical reinforcement learning algorithms, the goal is to compute or approximate V^π well. A generic way to approximate V^π is to produce a parametrized estimate $\hat{V}$ of V^π. Under linear function approximation, we model $\hat{V}$ as a linear approximation. To resolve the issue of computational tractability, we consider approximating V^π in a finite-dimensional subspace of $L^2(\mathcal{S}, \mu)$. Then $\hat{V}$ can be written as

$$\hat{V}(s, w) = \sum_{i=0}^{k} \phi_i(s) w_i \ , \quad \forall s \in \mathcal{S}$$

where $w = [w_1 \ w_2 \ \dots \ w_k]^\top \in \mathbb{R}^k$ is a parameter vector and $\Phi = \{\phi_1, \dots, \phi_k\}$ are a set of basis functions from $\mathcal{S}$ to $\mathbb{R}$ such that the span(Φ), is a finite-dimensional subset of $L^2(\mathcal{S}, \mu)$. We note that $\mathbb{R}^k$ can be viewed as the parameter space and span(Φ) as the approximation space. Typically in applications, Φ is chosen and fixed a-priori. Let us define $\phi(s) = \big(\phi_1(s), \phi_2(s), \dots, \phi_k(s)\big)^\top$. Then $\hat{V}$ can be represented compactly as an inner product

$$\hat{V}(s, w) = \langle \phi(s), w \rangle \ .$$

Example 3.1 (State Aggregation)**.** The following example is adapted from [3]. A special case of linear function approximation is known as state aggregation. In

state aggregation the state space $\mathcal{S}$ is partitioned into k disjoint sets $S_1, S_2, ..., S_k$. We then introduce a parameter vector $w \in \mathbb{R}^k$ and approximate each set S_l by a single element of the parameter vector as follows

$$\hat{V}^\pi(s) = w_l \quad \text{if } s \in S_l \ .$$

Thus, all states in the same subset S_l share the same parametrization and approximation of their value function. Now if we let:

$$\epsilon_l = \sup_{s_i, s_j \in S_l} \left| V^\pi(s_i) - V^\pi(s_j) \right| \quad l = 1, \ldots, k \ ,$$

so that ϵ_l represents the largest possible difference of true values of states in S_l, then picking w_l such that

$$w_l = \inf_{s_i \in S_l} V^\pi(s_i) + \frac{\epsilon_l}{2}$$

gives a maximum approximation error of

$$\sup_{s_i} \left| \hat{V}^\pi(s_i, w) - V^\pi(s_i) \right| = \max_l \frac{\epsilon_l}{2} \quad .$$

Thus V^π can be approximated well as long as the subsets S_l are chosen such that ϵ_l is small, that is V^π does not have substantial variation within each set.

When considering linear function approximation to generate approximate solutions, using the Bellman equations may not generate a representable solution. Consider the value iteration method applied to a linear function approximation $\hat{V}_k \in \text{span}(\Phi)$

$$\hat{V}_{k+1} \coloneqq \mathcal{T}\hat{V}_k \ .$$

Since the Bellman operator $\mathcal{T}$ is an affine linear operator, $\mathcal{T}\hat{V}_k$ may no longer be in $\text{span}(\Phi)$ and thus not be representable in the chosen approximation space. To resolve this issue, many methods consider projecting the entire Bellman equation

$$\hat{V} = \Pi \mathcal{T} \hat{V} \ ,$$

where Π is a projection operator with $\text{im}(\Pi) = \text{span}(\Phi)$. We will detail these methods in the next section.

3.2 Projected Equation Methods

In this section we introduce readers to the theory behind methods that look to approximate the value function by considering a projected form of the Bellman equations

$$\hat{V} = \Pi \mathcal{T} \hat{V} \ ,$$

where $\hat{V} \in \text{span}(\Phi)$ and Π is a projection operator with $\text{im}(\Pi) = \text{span}(\Phi)$. As an aside, the Galerkin method has a long history in computational mathematics for approximating higher dimensional equations via an approximate solution generated by the projected equations. As we will see in the chapter 4, the differentiating factor in approximate reinforcement learning comes from the introduction of Monte-Carlo simulation techniques.

3.2.1 Oblique Projection Operators

We consider the set of possible projection operators that can be applied to the Bellman equations to find an approximate solution. Projection operators that can project in *any* direction are collectively known as oblique projections. Under this general framework, we will provide a unified view to characterise the solution of all projected equation methods.

An oblique projection operator $\Pi : L^2(\mathcal{S}, \mu) \to L^2(\mathcal{S}, \mu)$ can be characterised by the the two subspaces $\text{im}(\Pi)$ and $\text{im}(\Pi^*)$ (see chapter 1). The purpose of looking at projection operators is to find approximations in lower-dimensional spaces that can approximate a given function well and be computationally tractable. Thus, we will focus on oblique projection operators with finite-dimensional image. The set of oblique projection operators we consider are then compact operators (see chapter 1). A key property of projection operators with finite dimensional image is that the image of the adjoint has the same dimension.

Proposition 3.2. *Let $\Pi : L^2(\mathcal{S}, \mu) \to L^2(\mathcal{S}, \mu)$ be a bounded linear operator with finite-dimensional image and let Π^* be its adjoint. Then the image of Π^* has the same dimension as the image of Π.*

Proof. Let $l_y(x) = \langle \Pi x, y \rangle_\mu$. Then by the Riesz representation theorem, there exists z_y such that

$$l_y(x) = \langle x, z_y \rangle_\mu \ ,$$

and by definition, $\Pi^* y = z_y$. Now note that if $y \in \text{im}(\Pi)^\perp$, then $y \in \ker(\Pi^*)$ (see preliminaries) and so $z_y = 0$. Now let $z \in L^2(\mathcal{S}, \mu)$. Since $L^2(\mathcal{S}, \mu)$ can be decomposed into the direct sum of $\text{im}(\Pi)$ and $\text{im}(\Pi)^\perp$, then there exists $z_1 \in \text{im}(\Pi)$ and $z_2 \in \text{im}(\Pi)^\perp$ such that

$$z = z_1 + z_2 \ .$$

Applying Π^* to z then gives

$$\Pi^* z = \Pi^* z_1 \ .$$

Thus since $z_1 \in \text{im}(\Pi)$, which is finite-dimensional, the image of Π^* must also have the same dimensions.

□

Given the above result, we define the set of projection operators that we are interested in for finding lower-dimensional approximations as follows.

Definition 3.3 (Natural Projection Operators)**.** Let $\Pi : L^2(\mathcal{S}, \mu) \to L^2(\mathcal{S}, \mu)$ be an oblique projection operator such that $\dim(\text{im}(\Pi)) = k$. Let $\Phi = \{\phi_1, ..., \phi_k\}$ be a basis for $\text{im}(\Pi)$. Let $\Psi = \{\psi_1, ..., \psi_k\}$ be a basis for $\text{im}(\Pi^*)$. Then Π is a compact operator and can be characterised by the two sets (Φ, Ψ). We call the set of all such operators the *natural projection operators*.

3.2.2 Characterising Projected Functions

The geometric ideas underlying the Galerkin method provide the main tools with which an approximation of a function found by projection can be characterised. The parameter vector found as the solution of a projected function can be seen as the solution to a system of linear equations. Suppose we are looking to approximate $V \in L^2(\mathcal{S}, \mu)$ by a function $\hat{V} \in \text{span}(\Phi)$. Then we can consider using a natural projection operator Π_Ψ characterised by (Φ, Ψ) to find $\hat{V}$ by

$$\hat{V} = \Pi_\Psi V \ .$$

Note that since $\hat{V} \in \text{span}(\Phi)$, it can be represented in a linear form as

$$\hat{V}(s) = \sum_{i=0}^{k} \phi_i(s) w_i \ , \quad \forall s \in \mathcal{S}$$

where $w_i \in \mathbb{R}$. Since the basis functions are known, the only task left is to find an expression for the parameter vectors. Since Π_Ψ projects orthogonally to $\text{im}(\Pi_\Psi^*)$ and onto $\text{im}(\Pi_\Psi)$, we have that $V - \hat{V} \perp_\mu \ker(\Pi_\Psi)$. Thus, for all $i = 1, \ldots, k$,

$$\langle \psi_i, V - \hat{V} \rangle_\mu = 0 \ .$$

Expanding $\hat{V}$,

$$\begin{aligned} 0 &= \left\langle \psi_i, V - \sum_{j=0}^{k} \phi_j w_j \right\rangle_\mu \\ &= \langle \psi_i, V \rangle_\mu - \sum_{j=0}^{k} \langle \psi_i, \phi_j \rangle_\mu w_j \ . \end{aligned}$$

Now re-arranging for w_j on the left gives

$$\sum_{j=0}^{k} \langle \psi_i, \phi_j \rangle_\mu w_j = \langle \psi_i, V \rangle_\mu \ .$$

If $G \in \mathbb{R}^{k \times k}$ and $r \in \mathbb{R}^k$ are defined by

$$\begin{aligned} G_{ij} &= \langle \psi_i, \phi_j \rangle_\mu \ , \quad i, \ j = 1, 2, ..., k \ , \\ r_i &= \langle \psi_i, V \rangle_\mu \ , \quad i = 1, 2, ..., k \end{aligned}$$

respectively then we have a system of k linear equations that can be written in vector-matrix form as

$$Gw = r \ .$$

Thus solving for the parameter w amounts to solving a system of linear equations. In the case where $\Phi = \Psi$, the matrix G is called the *Gram* matrix. As noted in [9], if Φ is an *orthogonal basis* for $\text{im}(\Pi_\Psi)$, the Gram matrix is a diagonal matrix and the solution to $Gw = r$ is simple:

$$w_i = \frac{r_i}{\langle \phi_i, \phi_i \rangle_\mu} \ , \quad i = 1, \ldots, k \ .$$

We now use these ideas to characterise solutions to projected Bellman equations.

3.2.3 The Projected Bellman Equations

Using the ideas developed in the last section, the solution to an naturally projected Bellman equation can be characterised by solving a system of linear equations as well. For notational convenience we first re-define elements of the Bellman

equation for easier exposition. For any $V \in \mathbb{R}^{\mathcal{S}}$, recall that the Bellman operator is given by

$$\mathcal{T}V(s) \coloneqq R(s) + \gamma \int_{\mathcal{S}} T(s'|s)V(s')ds' \ .$$

Let $P_T : L_2(\mathcal{S}, \mu) \to L_2(\mathcal{S}, \mu)$ be an operator defined as

$$P_T f(s) \coloneqq \int_{\mathcal{S}} T(s'|s)f(s')ds' \ .$$

We can thus express the Bellman operator as

$$\mathcal{T}V(s) \coloneqq R(s) + \gamma P_T V(s) \ .$$

As a further convenience, consider the Bellman equation applied to $\hat{V} \in \text{im}(\Phi)$ such that $\hat{V}(\cdot) = \phi^\top(\cdot)w$. Then we define P_T over a function returning a multi-dimensional vector to be taken component wise:

$$P_T\phi(s) \coloneqq \Big[P_T\phi_1(s) \dots P_T\phi_k(s)\Big]^\top \in \mathbb{R}^k \ .$$

We use the unconventional notation of applying the transpose superscript to a function to indicate taking the transpose of the resultant value:

$$P_T\phi^\top(s) \coloneqq \Big[P_T\phi_1(s) \dots P_T\phi_k(s)\Big] \ .$$

Now let Π_Ψ be an natural projection characterised by (Φ, Ψ). We look to solve the projected Bellman equation

$$\hat{V} = \Pi_\Psi \mathcal{T}\hat{V}$$

for $\hat{V} \in \text{span}(\Phi)$. If a solution exists, then there exists a parameter vector $w = (w_1, \dots, w_k)^\top \in \mathbb{R}^k$ such that

$$\hat{V}(\cdot) = \sum_{i=1}^{k} \phi_i(\cdot)w_i \ .$$

This parameter vector is characterised as the solution to a system of linear equations in the next section.

Proposition 3.4. *Let Π_Ψ be a natural projection operator characterised by (Φ, Ψ). Let $\hat{V} \in im(\Pi_\Psi)$ be given by $\hat{V} = \sum_{i=0}^{k} \phi_i w_i$. Consider the projected Bellman equation*

$$\hat{V} = \Pi_\Psi \mathcal{T}\hat{V} \ .$$

Let $G \in \mathbb{R}^{k\times k}$ and $r \in \mathbb{R}^k$ be defined by

$$G_{ij} = \langle \psi_i, (I - \gamma P_T)\phi_j \rangle_\mu \ , \quad i,j = 1,2,\ldots,k \ ,$$
$$r_i = \langle \psi_i, R \rangle_\mu \ , \quad i = 1,2,\ldots,k$$

respectively. Then if a solution to the projected Bellman equation exists, the parameter vector $w = (w_1, ..., w_k)^\top$ is given as the solution to the system of linear equations

$$Gw = r \ .$$

Proof. We now look to solve for w by performing the same derivation as in the previous section. We have for all $i = 1, ..., k$

$$\langle \psi_i, \mathcal{T}\hat{V} - \Pi_\Psi \mathcal{T}\hat{V} \rangle = 0 \ .$$

Substituting in the projected Bellman equation, we have:

$$\begin{aligned}
0 &= \langle \psi_i, \mathcal{T}\hat{V} - \hat{V} \rangle \\
&= \langle \psi_i, (R + \gamma P_T \phi^\top(\cdot)w) - \phi^\top(\cdot)w \rangle_\mu \\
&= \langle \psi_i, R \rangle_\mu + \left\langle \psi_i, \gamma \sum_{j=0}^{k} P_T \phi_j w_j \right\rangle_\mu - \left\langle \psi_i, \sum_{j=0}^{k} \phi_j w_j \right\rangle_\mu \\
&= \langle \psi_i, R \rangle_\mu - \sum_{j=0}^{k} \langle \psi_i, (\phi_j - \gamma P_T \phi_j) w_j \rangle_\mu \\
&= \langle \psi_i, R \rangle_\mu - \sum_{j=0}^{k} \langle \psi_i, (I - \gamma P_T)\phi_j w_j \rangle_\mu \ .
\end{aligned}$$

Re-arranging now gives

$$\sum_{j=0}^{k} \langle \psi_i, (I - \gamma P_T)\phi_j w_j \rangle_\mu = \langle \psi_i, R \rangle_\mu \ .$$

If $G \in \mathbb{R}^{k\times k}$ and $r \in \mathbb{R}^k$ are defined by

$$G_{ij} = \langle \psi_i, (I - \gamma P_T)\phi_j \rangle_\mu \ , \quad i,j = 1,2,\ldots,k \ ,$$
$$r_i = \langle \psi_i, R \rangle_\mu \ , \quad i = 1,2,\ldots,k$$

respectively then we have a linear system of k equations given by

$$Gw = r \ .$$

□

An immediate consequence is that we can define three key quantities that characterise the solution to any projected Bellman equation.

Corollary 3.5. *Let Π_Ψ be a natural projection operator characterised by (Φ, Ψ). Let $A \in \mathbb{R}^{k\times k}$, $B \in \mathbb{R}^{k\times k}$ and $b \in \mathbb{R}^k$ be defined by*

$$\begin{aligned} A_{ij} &= \langle \psi_i, \phi_j \rangle_\mu \ , \quad i,j = 1,2,\ldots,k \ , \\ B_{ij} &= \langle \psi_i, P_T \phi_j \rangle_\mu \ , \quad i,j = 1,2,\ldots,k \ , \\ b_i &= \langle \psi_i, R \rangle_\mu \ , \quad i = 1,2,\ldots,k \end{aligned}$$

respectively. If it exists, the unique fixed point of the projected Bellman equation given by

$$\hat{V} = \Pi_\Psi \mathcal{T} \hat{V}$$

depends only upon A, B, and b.

Proof. Given our definitions of A, B, and b, and the definitions of G and r in Proposition (3.4), we have that

$$\begin{aligned} G &= A - \gamma B \ , \\ b &= r \ . \end{aligned}$$

Thus, the linear system of equations solving for w can be expressed as

$$(A - \gamma B) w = b \ .$$

□

Our corollary suggests that *any* algorithm that solves for the fixed point of a projected Bellman equation has its solution characterised by the three quantities A, B, and b. We thus define our class of natural algorithms in relation to natural projection operators and make use of the property highlighted in Corollary 3.5 when we consider our results.

Definition 3.6 (Natural Algorithms)**.** Let Π_Ψ be a natural projection operator characterised by (Φ, Ψ). If an algorithm solves for the unique fixed point $\hat{V} \in \mathrm{im}(\Pi_\Psi)$ of the projected Bellman equation

$$\hat{V} = \Pi_\Psi \mathcal{T} \hat{V} \ ,$$

Then the algorithm is a *natural algorithm*. We say that such an algorithm is characterised by (Φ, Ψ).

In the next chapter we show that the general characterisation of projected Bellman equations and the formulation of natural algorithms still holds in the case of a finite state space. In the finite case, we will also see examples of natural algorithms, which will show that our class of natural algorithms is a non-empty and valid class of algorithms to consider for divergence results.

Chapter 4

Examples of Natural Algorithms

In this chapter we show that many common reinforcement learning methods are candidate natural algorithms in the case of a finite state space. Even though the state space is considered finite, we will assume that its size is 'large', that is the size of the state space presents computational difficulties for dynamic programming techniques and approximation methods are required. We begin by showing that the general characterisation of projected Bellman equations and the formulation of natural algorithms still holds in the case of a finite state space. We will then discuss the seminal methods in reinforcement learning, showing how they can be characterised as natural algorithms, before presenting counter-examples detailing divergence. Since we are looking to find sufficient conditions for divergence results, these counter-examples serve to illustrate key difficulties that plague natural algorithms even in the simpler (computational complexity wise) case of a finite state space.

4.1 The Projected Bellman Equations in Finite State Space

Before considering the projected Bellman equations we express for clarity the underlying value function space considered analogously to what was presented in chapter 3. When considering the case of a finite state space, it can be assumed for simplicity that $\mathcal{S} = \{1, \dots, n\}$. Thus the relevant value functions are all elements of $\mathbb{R}^n$. We assume that assumption (2) holds and that μ is the stationary distribution. We can then consider the inner product space induced by the inner product defined by the stationary distribution. Let $D_\mu \in \mathbb{R}^{n \times n}$ denote a diagonal

matrix given by

$$D_\mu = \begin{bmatrix} \mu(1) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \mu(n) \end{bmatrix} .$$

For any two vectors $x, y \in \mathbb{R}^n$, we define our inner product as

$$\langle x, y \rangle_\mu = x^\top D_\mu y \ .$$

Showing that $\langle \cdot, \cdot \rangle_\mu$ is an inner product is routine. The given inner product induces the norm $\|\cdot\|_\mu$ where $\|x\|_\mu = \sqrt{\langle x, x \rangle_\mu}$. We will analogously refer to $\|\cdot\|_\mu$ as the μ-weighted quadratic norm. We will similarly denote the set of vectors in $\mathbb{R}^n$ with finite μ-weighted quadratic norm by

$$L^2(\mathcal{S}, \mu) = \{V \in \mathbb{R}^n : \|V\|_\mu < \infty\} \ .$$

Again, note that for any stationary policy π, V^π is an element of $L^2(\mathcal{S}, \mu)$ since proposition 3.5 guarantees that V^π is bounded in the infinity norm.

Now consider oblique projection operators. Since $\mathbb{R}^n$ is finite dimensional, all oblique projection operators are trivially natural projection operators. Furthermore, since linear transformations are now given by matrices, we have that $\Pi \in \mathbb{R}^{n \times n}$ and the adjoint Π^* is equivalent to the transposed matrix $\Pi^\top$. We look to projection operators that project to a lower-dimensional subspace of $\mathbb{R}^n$. Now suppose we are given a natural projection Π_Ψ whose image has dimension $k < n$. Let $\{\phi_1, \dots, \phi_k\}$ be a basis spanning $\mathrm{im}(\Pi_\Psi)$ and let $\{\psi_1, \dots, \psi_k\}$ be a basis spanning $\mathrm{im}(\Pi_\Psi^\top)$. Since we are working with vectors, we can denote summarise these sets of basis vectors in matrix form

$$\Phi = \begin{bmatrix} | & & | \\ \phi_1 & \dots & \phi_k \\ | & & | \end{bmatrix} \in \mathbb{R}^{n \times k} \ , \qquad \Psi = \begin{bmatrix} | & & | \\ \psi_1 & \dots & \psi_k \\ | & & | \end{bmatrix} \in \mathbb{R}^{n \times k} \ .$$

Note that for a natural projection Π_Ψ characterised by (Φ, Ψ), any $\hat{V} \in \mathrm{im}(\Pi_\Psi)$ can be expressed as

$$\hat{V} = \Phi w$$

for some parameter vector $w \in \mathbb{R}^k$. We can now consider the solution to a projected Bellman equation of the form

$$\Phi w = \Pi_\Psi \mathcal{T} \Phi w \ .$$

The following proposition shows that w is uniquely determined as the solution to a linear system of equations when both Φ and Ψ are full rank.

Proposition 4.1. *Let Π_Ψ be a natural projection characterised by (Φ, Ψ). Then the projected Bellman equation*

$$\Phi w = \Pi_\Psi \mathcal{T} \Phi w$$

has the unique solution

$$w = \left(\Psi^\top D_\mu \left(I - \gamma T\right) \Phi\right)^{-1} \Psi^\top D_\mu R \ .$$

Proof. We first note that $\mathcal{T}\Phi w - \Pi_\Psi \mathcal{T}\Phi w \in \ker(\Pi)$. Since $\Pi_\Psi \mathcal{T}\Phi w \in \mathrm{im}(\Pi_\Psi)$, $\mathcal{T}\Phi w - \Pi_\Psi \mathcal{T}\Phi w \in \ker(\Pi)$ by the direct sum decomposition property of Π_Ψ. Thus we have that

$$\Psi^\top D_\mu \left(\mathcal{T}\Phi w - \Pi_\Psi \mathcal{T}\Phi w\right) = 0 \ .$$

Substituting in $\Pi_\Psi \mathcal{T}\Phi w = \Phi w$ and the definition of $\mathcal{T}$ gives

$$\Psi^\top D_\mu \left(R + \gamma T \Phi w - \Phi w\right) = 0$$

which after re-arranging gives

$$\Psi^\top D_\mu \left(I - \gamma T\right) \Phi w = \Psi^\top D_\mu R$$

Note that $I - \gamma T$ is always invertible. To see this, note that the absolute value of any eigenvalues of a stochastic matrix is less than or equal to 1 (see chapter 1); hence $I - \gamma T$ has positive eigenvalues greater than 0. Then since Ψ and Φ are full rank and D_μ is a diagonal matrix, $\Psi^\top D_\mu \left(I - \gamma T\right) \Phi$ is invertible. Thus we have that the unique solution to the generalised Bellman equation is given by

$$\hat{w} = \left(\Psi^\top D_\mu \left(I - \gamma T\right) \Phi\right)^{-1} \Psi^\top D_\mu R \ .$$

□

Again, we can immediately as a consequence show that the solution to any projected Bellman equation depends on three key quantities.

Corollary 4.2. *Let $A = \Psi^\top D_\mu \Phi$, $B = \Psi^\top D_\mu T \Phi$, and $b = \Psi^\top D_\mu R$. Then the solution to any projected Bellman equation depends only upon A, B and b.*

Proof. From proposition 4.1, we have that the unique solution to the projected Bellman equation is given by

$$\hat{w} = \left(\Psi^\top D_\mu \left(I - \gamma T\right) \Phi\right)^{-1} \Psi^\top D_\mu R \ .$$

Substituting for A, B, and b gives the desired result

$$\hat{w} = \left(A - \gamma B\right)^{-1} b \ . \tag{4.1}$$

□

Thus we can clearly see that by re-defining Φ and Ψ as matrices, our definition of natural algorithms still carries through in the finite case. We will now proceed to survey some important reinforcement learning techniques and the different ideas that underpin their usage. Along the way we will show that these algorithms can all be classified as natural algorithms.

4.2 Simulation-Based Methods

In this section we look to methods that resolve reinforcement learning problems in the case where the state transition dynamics of the environment are unknown but the resultant state sequence can be simulated. Under guarantees of ergodicity in the underlying state Markov process, statistical estimates of key quantities in the projected Bellman equations are guaranteed to converge to the correct quantities by the law of large numbers. Indeed, the validity of Monte-Carlo methods arises as a direct consequence of the law of large numbers. As such, we will not present any description of Monte-Carlo methods here and leave its illustration to an example in the appendix [see appendix A]. Note that the results presented in this section are adapted from [2] for our usage.

Recall that the solution to a projected Bellman equation is given by

$$(A - \gamma B)w = b \ ,$$

where

$$A = \Phi^\top D_\mu \Phi \ , B = \Phi^\top D_\mu T \Phi \ , b = \Phi^\top D_\mu R \ .$$

The methods in this section primarily find prominence when T and μ are unknown. In this case, the simple solution is to consider producing statistical estimates of A, B, and b. Consider the following. Suppose we simulate an infinitely

long sequence of the state Markov process $s_0, s_1, s_2, \ldots$ with stationary distribution μ. At time t, we compute $\phi(s_t)$ and $R(s_t)$, and after generating the transition (s_t, s_{t+1}) we form $\phi(s_{t+1})$. One particularly natural set of estimates to compute is

$$\begin{aligned} A_t &\coloneqq \frac{1}{t+1} \sum_{k=0}^{t} \phi(s_k)\phi(s_k)^\top \ , \\ B_t &\coloneqq \frac{1}{t+1} \sum_{k=0}^{t} \phi(s_k)\phi(s_{k+1})^\top \ , \\ b_t &\coloneqq \frac{1}{t+1} \sum_{k=0}^{t} \phi(s_k)R(s_k) \end{aligned}$$

for all $n \in \mathbb{N}$ and where $\phi(s) = \big(\phi_1(s), \ldots, \phi_k(s)\big)^\top$. The next proposition shows that A_n, B_n and b_n converge to A, B, and b.

Proposition 4.3. *Let*

$$\begin{aligned} A_t &\coloneqq \frac{1}{t+1} \sum_{k=0}^{t} \phi(s_k)\phi(s_k)^\top \ , \\ B_t &\coloneqq \frac{1}{t+1} \sum_{k=0}^{t} \phi(s_k)\phi(s_{k+1})^\top \ , \\ b_t &\coloneqq \frac{1}{t+1} \sum_{k=0}^{t} \phi(s_k)R(s_k) \end{aligned}$$

for all $t \in \mathbb{N}$ *and where* $\phi(s) = \big(\phi_1(s), \ldots, \phi_k(s)\big)^\top$. *Then* $(A_t, B_t, b_t) \xrightarrow{t\to\infty} (A, B, b)$.

Proof. The proof here is an adaptation of a similar derivation in [2]. We first note that A, B and b can be written as

$$\begin{aligned} A &= \sum_{s=1}^{n} \mu(s)\phi(s)\phi(s)^\top \ , \\ B &= \sum_{s=1}^{n} \mu(s)\phi(s) \sum_{s'=1}^{n} T(s'|s)\phi(s')^\top \ , \\ b &= \sum_{s=1}^{n} \mu(s)\phi(s)R(s) \ . \end{aligned}$$

Now consider A_t. We have the following derivation:

$$\begin{aligned} A_t &\coloneqq \frac{1}{t+1}\sum_{k=0}^{t}\phi(s_k)\phi(s_k^\top) \\ &= \sum_{k=0}^{t}\sum_{s=1}^{n}\frac{\mathbf{1}(s_k=s)}{t+1}\phi(s)\phi(s)^\top \\ &= \sum_{s=1}^{n}\frac{\sum_{k=0}^{t}\mathbf{1}(s_k=s)}{t+1}\phi(s)\phi(s)^\top \\ &= \sum_{s=1}^{n}\hat{\mu}(s)\phi(s)\phi(s)^\top \end{aligned}$$

where $\hat{\mu}(s) \coloneqq \frac{\sum_{k=0}^{t}\mathbf{1}(s_k=s)}{t+1}$. By the law of large numbers we have that $\hat{\mu}(s) \xrightarrow{t\to\infty} \mu(s)$. Thus, $A_t \xrightarrow{t\to\infty} A$. For B_t we have a similar derivation,

$$\begin{aligned} B_t &\coloneqq \frac{1}{t+1}\sum_{k=0}^{t}\phi(s_k)\phi(s_{k+1})^\top \\ &= \frac{1}{t+1}\sum_{k=0}^{t}\sum_{s=1}^{n}\sum_{s'=1}^{n}\mathbf{1}(s_k=s,s_{k+1}=s')\phi(s)\phi(s')^\top \\ &= \sum_{s=1}^{n}\sum_{s'=1}^{n}\frac{\sum_{k=0}^{t}\mathbf{1}(s_k=s,s_{k+1}=s')}{t+1}\phi(s)\phi(s')^\top \\ &= \sum_{s=1}^{n}\sum_{s'=1}^{n}\frac{\sum_{k=0}^{t}\mathbf{1}(s_k=s)\cdot\sum_{k=0}^{t}\mathbf{1}(s_k=s,s_{k+1}=s')}{t+1\cdot\sum_{k=0}^{t}\mathbf{1}(s_k=s)}\phi(s)\phi(s')^\top \\ &= \sum_{s=1}^{n}\frac{\sum_{k=0}^{t}\mathbf{1}(s_k=s)}{t+1}\phi(s)\left(\sum_{s'=1}^{n}\frac{\sum_{k=0}^{t}\mathbf{1}(s_k=s,s_{k+1}=s')}{\sum_{k=0}^{t}\mathbf{1}(s_k=s)}\phi(s')^\top\right) \\ &= \sum_{s=1}^{n}\hat{\mu}(s)\phi(s)\sum_{s'=1}^{n}\hat{T}(s'|s)\phi(s')^\top \end{aligned}$$

where $\hat{T}(s'|s) = \frac{\sum_{k=0}^{t}\mathbf{1}(s_k=s,s_{k+1}=s')}{\sum_{k=0}^{t}\mathbf{1}(s_k=s)}$. Again, by the law of large numbers, we have that $\hat{T}(s'|s) \xrightarrow{t\to\infty} T(s'|s)$. Thus, $B_t \xrightarrow{t\to\infty} B$. We can similarly write b_t as

$$\begin{aligned} b_t &\coloneqq \frac{1}{t+1}\sum_{k=0}^{t}\phi(s_k)R(s_k) \\ &= \sum_{s=0}^{n}\frac{\sum_{k=0}^{t}\mathbf{1}(s_k=s)}{t+1}\phi(s)R(s) \\ &= \sum_{s=0}^{n}\hat{\mu}(s)\phi(s)R(s)\ . \end{aligned}$$

Thus, $b_t \xrightarrow{t\to\infty} b$. □

Example 4.4. The least-squares temporal difference method (LSTD), first proposed by Bradke and Barto [5], is one such reinforcement learning method that approximately solves the projected Bellman equation by directly evaluating A_t, B_t, and b_t. To find an approximate solution, the LSTD algorithm directly evaluates

$$(A_t - \gamma B_t)\,\hat{w}_t = b_t \tag{4.2}$$

to solve for $\hat{w}_t$. A unique solution at each time step is guaranteed to exist so long as $A_t - \gamma B_t$ is invertible. We also note that by writing equation (4.2) as

$$(A_t - \gamma B_t)\,\hat{w}_t - b_t = 0\ ,$$

we have that

$$\begin{aligned} 0 &= \frac{1}{t+1}\left(\sum_{k=0}^{t}\phi(s_k)\phi(s_k)^\top - \gamma\sum_{k=0}^{t}\phi(s_k)\phi(s_{k+1})^\top\right)\hat{w}_k - \frac{1}{t+1}\sum_{k=0}^{t}\phi(s_k)R(s_k) \\ &= \sum_{k=0}^{t}\phi(s_k)\left(\phi(s_k)^\top\hat{w}_k - \gamma\phi(s_{k+1})^\top\hat{w}_t - R(s_k)\right) \qquad (4.3) \\ &= \sum_{k=0}^{t}\phi(s_k)d_k\ , \end{aligned}$$

where $d_k \coloneqq \phi(s_k)^\top\hat{w}_k - \gamma\phi(s_{k+1})^\top\hat{w}_t - R(s_k)$ is known as the *temporal difference*. Noting that $(A - \gamma B)\,\hat{w}_k - b$ can be written as

$$(A - \gamma B)\,w - b = \Phi^\top D_\mu\left((\Phi - \gamma T\Phi)\,w - R\right)\ ,$$

we see that $\phi(s_k)d_t$ can be seen as a sample of $(A - \gamma B)\,w - b$ and that the LSTD method in essence looks for $\hat{w}$ that minimises the sum over all temporal differences up to time t. Thus LSTD is clearly a natural algorithm since it looks to directly estimate A, B and b and produces a solution based off these three quantities in the limit.

Remark 4.5. We note that since LSTD method computes the feature vector of each state at time t, $\phi(s_t)$, the algorithm uses the sequence of *feature vectors* generated, that is $\{\phi(s_t)\}_{t\in\mathbb{N}}$, rather than the state sequence $\{s_t\}_{t\in\mathbb{N}}$ to perform its updates. This is a common theme in methods that combine sampling with function approximation.

4.2.1 Temporal-Difference Learning

A seminal idea in the field of reinforcement learning is undoubtedly temporal difference learning. Temporal difference methods, first proposed by Sutton [16], are a family of sample-based methods that incrementally update the parameter vector to approximate the value function. We will present the main algorithm before focusing in on a special case known as the $TD(0)$ method.

The $TD(\lambda)$ Algorithm

Recall that the value function is given by

$$V^\pi(s) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t R^\pi(S_t) \middle| S_0 = s\right]$$

where $\gamma \in (0,1)$ is the discount factor. Under a function approximation setting, we look to find an approximation function $\hat{V}(\cdot, w)$ parametrized by w to approximate V^π. The *temporal difference* at time step t is given by

$$d_t = R^\pi(s_t) + \gamma \hat{V}^\pi(s_{t+1}, w_t) - \hat{V}^\pi(s_t, w_t).$$

Then for $\lambda < 1$, the $TD(\lambda)$ algorithm updates the parameter vector w by the update rule

$$w_{t+1} \coloneqq w_t + \alpha_t d_t \sum_{k=0}^{t} (\gamma\lambda)^{t-k} \nabla_w \hat{V}(s_t, w_t)$$

where $\nabla_w \hat{V}$ denotes the vector of $\hat{V}$'s partial derivatives. The $TD(\lambda)$ algorithm really defines a class of algorithms since each value of λ parametrizes the algorithm's weighting of future rewards. To see this, note that for $\lambda = 0$, the $TD(0)$ update is given by $w_{t+1} = w_t + \alpha_t d_t \nabla_w \hat{V}(s_t, w_t)$, meaning only one-step samples are considered, whereas for $\lambda = 1$, the $TD(1)$ update considers the entire trajectory.

Now suppose that linear function approximation is used. Then $\hat{V} = \Phi w \in \mathrm{im}(\Phi)$ where $\Phi = [\phi_1 | \ldots | \phi_k]$ is an $n \times k$ matrix with basis functions $\phi_1, \phi_2, ..., \phi_k \in \mathbb{R}^n$ forming its columns. Let $\phi(s) = (\phi_1(s), \phi_2(s), ..., \phi_k(s))^\top$. Thus the gradient of $\hat{V}$ is given by $\nabla_w \hat{V}(s) = \phi(s)$. The $TD(\lambda)$ update rule can now be explicitly expressed as

$$w_{t+1} = w_t + \alpha_t d_t \sum_{k=0}^{t} (\gamma\lambda)^{t-k} \phi(s_k) \quad .$$

To simplify the notation, we define the *eligibility vector* at time t to be

$$z_t = \sum_{k=0}^{t} (\gamma\lambda)^{t-k}\phi(i_k)$$

which can also be updated incrementally by $z_{t+1} = \gamma\lambda z_t + \phi(s_{t+1})$. Thus, the $TD(\lambda)$ update rule is more compactly given by

$$w_{t+1} = w_t + \alpha_t d_t \sum_{k=0}^{t} (\gamma\lambda)^{t-k}\phi(s_k) \quad .$$

In what follows, we focus exclusively on the $TD(0)$ algorithm.

The Convergence of $TD(0)$

In this section we show that the $TD(0)$ algorithm with linear function approximation converges to the fixed point of the orthogonally projected Bellman equation, thus showing that $TD(0)$ is another algorithm within the class of natural algorithms. To begin with, we first re-iterate key assumptions.

Assumption 3. Assume that:

- The Markov chain $\{S_t\}_{t\in\mathbb{N}}$ is ergodic and that there exists a stationary distribution μ.
- The matrix Φ with $\{\phi_1, \ldots, \phi_k\}$ as its columns is full rank.

Given assumption 3, the orthogonal projection is given by

$$\Pi_\Phi = \Phi \left(\Phi^\top D_\mu \Phi\right)^{-1} \Phi^\top D_\mu \ ,$$

and the orthogonally projected Bellman equation is given by

$$\Phi\hat{w} = \Pi_\Phi \mathcal{T} \Phi\hat{w} \ .$$

For this equation to have a solution, $\Pi_\Phi\mathcal{T}$ must be a contraction mapping with respect to $\|\cdot\|_\mu$. To show this we will show that Π_Φ is a non-expansion and that $\mathcal{T}$ is a contraction mapping with respect to $\|\cdot\|_\mu$. The next result, shown in [21], says that the transition matrix is a non-expansion.

Lemma 4.6. *[21] The transition matrix T is a non-expansion with respect to $\|\cdot\|_\mu$, that is for all $V \in L^2(\mathcal{S}, \mu)$, $\|TV\|_\mu \leq \|V\|_\mu$.*

Proof. We have the following derivation

$$\begin{aligned}
\|TV\|_\mu^2 &= V^\top T^\top D_\mu TV \\
&= \sum_{s=1}^{n} \mu(s) \left(\sum_{s'=1}^{n} T(s'|s) V(s') \right)^2 \\
&\overset{(a)}{\leq} \sum_{s=1}^{n} \mu(s) \sum_{s'=1}^{n} T(s'|s) (V(s'))^2 \\
&\overset{(b)}{=} \sum_{s'=1}^{n} \sum_{s=1}^{n} \mu(s) T(s'|s) (V(s'))^2 \\
&\overset{(c)}{=} \sum_{s'=1}^{n} \mu(s') (V(s'))^2 \\
&= \|V\|_\mu^2 \quad .
\end{aligned}$$

(a) follows by Jensen's inequality, (b) follows by the Tonelli-Fubini theorem and finally, (c) follows since μ is the stationary distribution. Our derived inequality implies that $\|TV\|_\mu \leq \|V\|_\mu$ holds since the quadratic function is monotonic increasing on $\mathbb{R}_+$. □

The next result establishes that orthogonal projections are also non-expansions.

Lemma 4.7. *[21]. Let Π_Φ be an orthogonal projection. Then Π is non-expansive, i.e.*

$$\left\| \Pi_\Phi V - \Pi_\phi \tilde{V} \right\|_\mu \leq \left\| V - \tilde{V} \right\|_\mu \ , \forall V, \tilde{V} \in \mathbb{R}^n \ .$$

Proof. By the linearity of Π_Φ, we have

$$\begin{aligned}
\left\| \Pi_\Phi V - \Pi_\Phi \tilde{V} \right\|_\mu^2 &= \left\| \Pi_\Phi (V - \tilde{V}) \right\|_\mu^2 \\
&\leq \left\| \Pi_\Phi (V - \tilde{V}) \right\|_\mu^2 + \left\| (I - \Pi_\Phi)(V - \tilde{V}) \right\|_\mu^2 \ . \qquad (*)
\end{aligned}$$

For any vector $x \in \mathbb{R}^n$,

$$x = \Pi_\Phi(x) + (I - \Pi_\Phi)(x) \ .$$

Thus, applying Π_Φ to both sides gives

$$\Pi_\Phi(x) = \Pi_\Phi^2(x) + \Pi_\Phi(I - \Pi_\Phi)(x) = \Pi_\Phi(x) + \Pi_\Phi(I - \Pi_\Phi)(x)$$

since projection operators are idempotent. Thus it must be the case that $\Pi_\Phi(I - \Pi_\Phi)(x) = 0$, implying that $(I - \Pi_\Phi)(x) \in \ker(\Pi_\Phi)$ and is thus orthogonal to $\mathrm{im}(\Pi)$. Thus by the Pythagorean theorem we have

$$\left\|\Pi(V - \tilde{V})\right\|_\mu^2 + \left\|(I - \Pi)(V - \tilde{V})\right\|_\mu^2 = \left\|V - \tilde{V}\right\|_\mu^2 .$$

Composing the above inequality with (*) shows the desired property. □

Given these two results, we have the following proposition stating that $\mathcal{T}$ and $\Pi\mathcal{T}$ are contraction mappings with respect to $\|\cdot\|_\mu$.

Proposition 4.8. *[21]. The operators $\mathcal{T}$ and $\Pi\mathcal{T}$ are contraction mappings with respect to $\|\cdot\|_\mu$.*

Proof. We first note that for a finite state space, $\mathcal{T}$ can be expressed as $\mathcal{T}V = R + \gamma TV$. Thus, for any $V, \tilde{V} \in \mathbb{R}^n$

$$\begin{aligned}\left\|\mathcal{T}V - \mathcal{T}\tilde{V}\right\|_\mu &\overset{(a)}{=} \left\|\gamma TV - \gamma T\tilde{V}\right\|_\mu \\ &\overset{(b)}{=} \gamma\left\|T(V - \tilde{V})\right\|_\mu \\ &\overset{(c)}{\leq} \gamma\left\|V - \tilde{V}\right\|_\mu .\end{aligned}$$

where (a) follows by definition, (b) by re-arranging, and (c) since T is a non-expansion (Lemma 4.6). This establishes that $\mathcal{T}$ is a contraction w.r.t. $\|\cdot\|_\mu$. Now since Π is a non-expansion (Lemma 4.7), we also have that $\Pi\mathcal{T}$ is a contraction mapping. □

Thus, by Banach's fixed point theorem, the orthogonally projected Bellman equation

$$\Phi\hat{w} = \Pi_\Phi\mathcal{T}\Phi\hat{w}$$

is guaranteed to have a unique solution. By proposition 4.1, we have that $\hat{w}$ is uniquely determined by

$$\hat{w} = \left(\Phi^\top D_\mu \left(I - \gamma T\right)\Phi\right)^{-1}\Phi^\top D_\mu R .$$

The seminal work by Tsitsiklis and Van Roy in [21] goes on to show that the $TD(0)$ algorithm converges to the fixed point of equation (4.3.2). A key assumption is that the realised state sequence $\{S_t\}_{t\in\mathbb{N}}$ that $TD(0)$ samples from to perform updates is in its steady state at all $t \in \mathbb{N}$. Indeed, as we will see in a later section, in many famous counter-examples, this condition is violated. Nevertheless, given that $TD(0)$ has been provably shown to compute the unique fixed point of 4.3.2, it is also a natural algorithm.

4.3 Bellman Error Methods

In this section we take a look at another class of common methods known as Bellman error methods. Bellman error methods look to minimise the difference in the two sides of the Bellman fixed point equation directly to find a parameter vector. Bellman error methods amount to minimising

$$\min_w \|\Phi\hat{w} - \mathcal{T}\Phi\hat{w}\|_\mu^2$$

for the parameter w. The error function is known as the *Bellman error*. Expanding the Bellman operator and re-arranging gives the simpler form

$$\min_w \left\|(I - \gamma T)\,\Phi\hat{w} - R\right\|_\mu^2 \ .$$

Since $\|\cdot\|_\mu^2$ is a coercive convex function, setting its gradient to 0 gives a necessary and sufficient condition for the unique global minimiser:

$$\begin{aligned} 0 &= \nabla_w \|\Phi\hat{w} - \mathcal{T}\Phi\hat{w}\|_\mu^2 \\ &\overset{(a)}{=} \left((I - \gamma T)\,\Phi\right)^\top D_\mu \left(\Phi\hat{w} - (R + \gamma T\Phi\hat{w})\right) \\ &= \left((I - \gamma T)\,\Phi\right)^\top D_\mu \left((I - \gamma T)\Phi\hat{w} - R\right) \ . \end{aligned}$$

Here (a) follows by expanding the Bellman operator and then taking the gradient w.r.t. w. Re-arranging for w gives the unique solution

$$\hat{w} = \left(\left((I - \gamma T)\,\Phi\right)^\top D_\mu (I - \gamma T)\Phi\right)^{-1} \left((I - \gamma T)\,\Phi\right)^\top D_\mu R$$

where we note that $\left((I - \gamma T)\,\Phi\right)^\top D_\mu (I - \gamma T)\Phi$ is invertible since $(I - \gamma T)$ is invertible and Φ is full rank (for full justification, see proposition 4.1).

4.3.1 The Residual Gradient Method

The residual gradient method, first proposed by Baird [1], is an example of a reinforcement learning algorithm that looks to minimise the Bellman error. In essence, the algorithm is a gradient method that performs the method of steepest descent on the Bellman error function. For an unconstrained optimisation problem with a differentiable objective function $f : \mathbb{R}^n \to \mathbb{R}$, the method of steepest descent generates a sequence of values by the update rule

$$x_{k+1} = x_k + \alpha_k \nabla_x f(x) \ ,$$

where α_k is the step-size. The step-size can be chosen in many different ways. To select α_k optimally, i.e. in the sense that

$$f(x_k + \alpha_k \nabla_x f(x)) = \min_{\alpha > 0} f(x_k + \alpha \nabla_x f(x)) \ ,$$

then α_k must satisfy

$$\frac{d}{d\alpha} f(x_k + \alpha \nabla f(x_k))\bigg|_{\alpha=\alpha_k} = 0 \ .$$

Thus, α_k must equivalently satisfy

$$\nabla f(x_k)^\top \nabla f(x_k + \alpha_k \nabla f(x_k)) = 0 \ .$$

The following theorem establishes a relevant case of convergence for the steepest descent method.

Theorem 4.9. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable, coercive convex function. Then the sequence $\{x_k\}$ generated by the method of steepest descent from any initial point x_0 converges the unique global minimiser of f.*

Proof. See [12]. □

Letting $f(w) = \|\Phi\hat{w} - \mathcal{T}\Phi\hat{w}\|_\mu^2$, the Residual Gradient algorithm's update rule is then given by

$$\begin{aligned} w_{k+1} &= w_k + \alpha_k \nabla_w f(w) \\ &= w_k + \alpha_k \left((I - \gamma T)\,\Phi\right)^\top D_\mu \left((I - \gamma T)\Phi w_k - R\right) \ , \end{aligned}$$

Which we can see by theorem 4.9 is guaranteed to converge to the unique global minimiser of f, which we recall is given by

$$\hat{w} = \left(\left((I - \gamma T)\,\Phi\right)^\top D_\mu (I - \gamma T)\Phi\right)^{-1} \left((I - \gamma T)\,\Phi\right)^\top D_\mu R \ .$$

4.3.2 A Unified Perspective

A key motivating factor in the research investigated in this thesis is the fact that both the Bellman error and projected Bellman equation methods can be unified into one framework using oblique projection operators.

Recall in section 3.3.3 that we showed that the $TD(0)$ algorithm minimises

$$E_{TD} = \|\Phi w^* - \Pi_\Phi \mathcal{T}\Phi w^*\|_\mu^2 \ ,$$

where Π_Φ is the orthogonal projection and the unique solution is given by

$$w_{TD} = \left(\Phi^\top D_\mu \left(I - \gamma T\right) \Phi\right)^{-1} \Phi^\top D_\mu R \; .$$

Similarly, the Residual Gradient algorithm minimises

$$E_{BE} = \|\Phi w^* - \mathcal{T}\Phi w^*\|_\mu^2$$

with the unique solution being given by

$$w_{BE} = \left(\left((I - \gamma T) \Phi\right)^\top D_\mu (I - \gamma T)\Phi\right)^{-1} \left((I - \gamma T) \Phi\right)^\top D_\mu R \; .$$

To simplify notation, let $L \coloneqq I - \gamma T$. Then w_{TD} and w_{BE} become

$$w_{TD} = \left(\Phi^\top D_\mu L \Phi\right)^{-1} \Phi^\top D_\mu R \quad \text{and} \quad w_{BE} = \left(\Phi^\top L^\top D_\mu L \Phi\right)^{-1} (L\Phi)^\top D_\mu R \; .$$

The following result by Scherrer [15] specifies the relationship between the two error methods.

Proposition 4.10. *[15]. The Bellman error is an upper bound of the $TD(0)$ error with respect to the μ-weighted quadratic norm.*

Proof. Let $\hat{V} = \Phi w \in \text{im}(\Phi)$. Since Π_Φ is an orthogonal projection w.r.t. the μ-weighted quadratic norm onto $\text{im}(\Phi)$, $\mathcal{T}\hat{V} - \Pi_\Phi \mathcal{T}\hat{V} \in \ker(\Phi)$. Also, since $\hat{V} \in \text{im}(\Phi)$, $\hat{V} - \Pi_\Phi \mathcal{T}\hat{V} \in \text{im}(\Phi)$. Then by the the Pythagorean theorem,

$$\left\|\hat{V} - \mathcal{T}\hat{V}\right\|_\mu^2 = \left\|\hat{V} - \Pi_\Phi \mathcal{T}\hat{V}\right\|_\mu^2 + \left\|\mathcal{T}\hat{V} - \Pi_\Phi \mathcal{T}\hat{V}\right\|_\mu^2 \; .$$

Thus, the Bellman error, $\left\|\hat{V} - \mathcal{T}\hat{V}\right\|_\mu^2$ upper bounds the TD(0) error $\left\|\hat{V} - \Pi_\Phi \mathcal{T}\hat{V}\right\|_\mu^2$. □

The term $\left\|\mathcal{T}\hat{V} - \Pi_\Phi \mathcal{T}\hat{V}\right\|_\mu^2$ can be interpreted as a measure of the orthogonal projection's adequacy w.r.t. the bellman operator $\mathcal{T}$. An important point here is that even if $TD(0)$ converges to the minimiser of E_{TD}, the solution may not be optimal since the projection may not be adequate.

Algorithms characterized by these two error functions can be seen as opposing extremes for methods whose solution is characterised by a projection on the bellman equations: bellman error methods have solutions characterised by taking the identity projection whereas $TD(0)$'s solution is characterised by the closest approximation to the bellman equation estimate, which is given by the orthogonal projection. The following proposition is adapted from [15] and unifies both methods into a single obliquely projected Bellman equation framework.

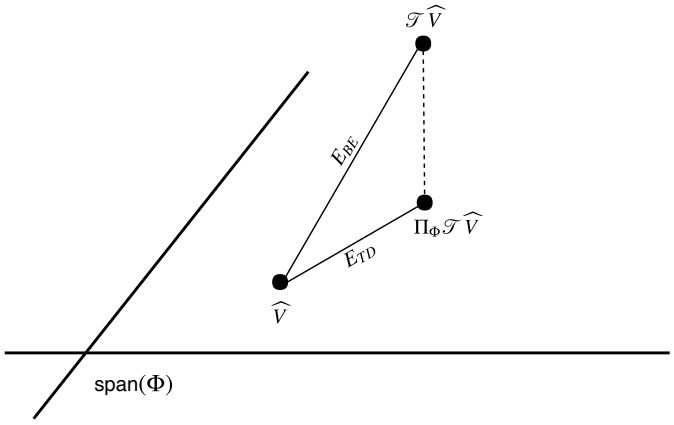

Figure 4.1: An illustration of the geometry and the corresponding Bellman error and TD error. Note here that the curly T corresponds to the Bellman operator.

Proposition 4.11. *[15].*
Let $X_{TD} = \Phi$ and $X_{BE} = (I - \gamma T)\,\Phi$. Let (Φ, X) characterise the projection matrix Π_X: Π_X has range span(Φ) and projects orthogonally to the subspace span(X). Π_X is given by:

$$\Pi_X = \Phi(X^\top D_\mu \Phi)^{-1} X^\top D_\mu$$

Then assuming that the transition matrix T of the underlying MDP is non-singular, methods that minimise the TD(0) error and Bellman error, with $X = X_{TD}$ and $X = X_{BE}$ respectively, solve the projected Bellman equation

$$\Phi \hat{w}_X = \Pi_X \mathcal{T} \Phi \hat{w}_X \ .$$

Proof. Let $\pi_X = \left(X^\top D_\mu \Phi\right)^{-1} X^\top D_\mu$. Then we note that

$$\begin{aligned} \pi_X \Pi_X &= \left(X^\top D_\mu \Phi\right)^{-1} X^\top D_\mu \Phi (X^\top D_\mu \Phi)^{-1} X^\top D_\mu \\ &= \pi_X \ . \end{aligned}$$

Then multiplying the Π_X projected Bellman equation by π_X gives

$$\begin{aligned} \hat{w}_X &= \pi_X \Pi_X \mathcal{T} \Phi \hat{w}_X \\ &= \pi_X \left(R + \gamma T \Phi \hat{w}_X\right) \\ &= \left(X^\top D_\mu \Phi\right)^{-1} X^\top D_\mu \left(R + \gamma T \Phi \hat{w}_X\right) \ . \end{aligned}$$

Re-arranging for $\hat{w}_X$ we have

$$\left(I - \gamma (X^\top D_\mu \Phi)^{-1} X^\top D_\mu T \Phi\right) \hat{w}_X = (X^\top D_\mu \Phi)^{-1} X^\top D_\mu R$$

Multiplying both sides by $X^\top D_\mu \Phi$,

$$X^\top D_\mu \left(I - \gamma T\right) \Phi \hat{w}_X = X^\top D_\mu R$$

Now note since $(I - \gamma T)$ has non-negative eigen-values (see proof of proposition 4.1) and Φ is full rank, $X^\top D_\mu \left(I - \gamma T\right) \Phi$ is invertible for both $X = X_{TD}$ and $X = X_{BE}$. Thus $\hat{w}_X$ is uniquely given by

$$\hat{w}_X = \left(X^\top D_\mu \left(I - \gamma T\right) \Phi\right)^{-1} X^\top D_\mu R \ .$$

Substituting in either X_{TD} or X_{BE} for X shows that $\hat{w}_{X_{TD}} = w_{TD}$ and $\hat{w}_{X_{BE}} = w_{BE}$. □

This result suggests that there is a spectrum of reinforcement learning algorithms that can all be characterised as solving obliquely projected Bellman equations.

4.4 Examples of Divergence

In this section we present some well-known divergence results that occur when temporal difference learning is combined with linear function approximation. In particular we will consider the $TD(0)$ algorithm again with one-step samples generated according to some initial distribution. The $TD(0)$ update in linear function approximation is given by

$$w_{t+1} = w_t + \alpha_t \big(R + \gamma \hat{V}(s_{t+1}) - V(s_t)\big)\phi(s_t)$$

The two results we will consider are Baird's counter-example and Tsitsiklis and Van Roy's counter-example.

4.4.1 Baird's Counter-example

Baird's counter-example [1] presents a 6 state MDP where each state transitions to the sixth state with probability 1. Consider the value function estimate under linear function approximation as shown for each state in the graph. For example, the value function estimate at state 1 is given by $\hat{V}(1) = w_0 + 2w_1 = [1\ 2\ 0\ 0\ 0\ 0\ 0]w$. Note that the linear function approximation architecture is chosen to have seven basis functions. The reward at all states on all actions is 0. Thus the optimal

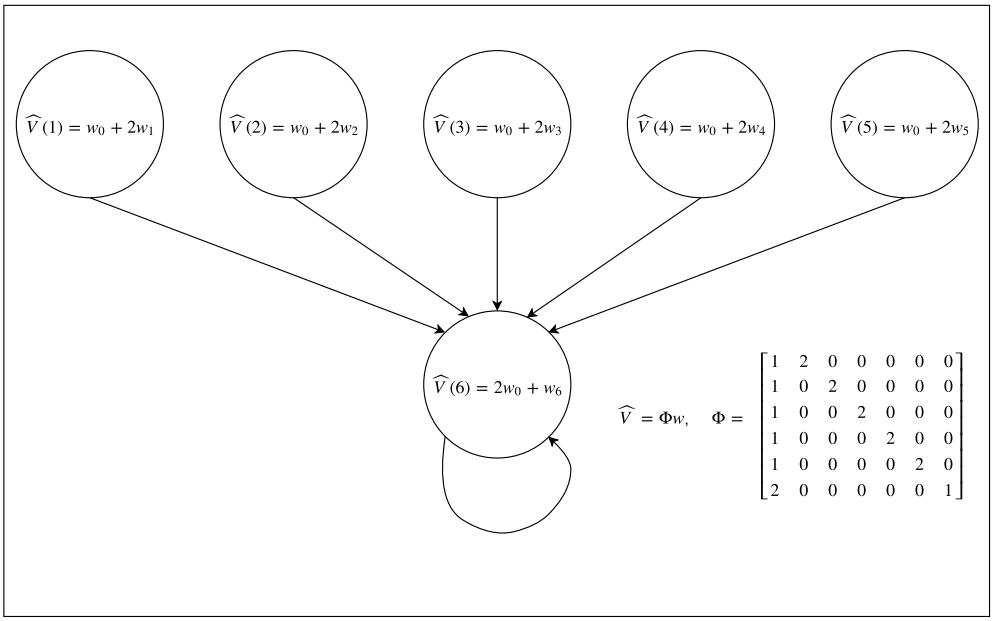

Figure 4.2: Baird's counter-example.

value function V^* is given by $V^* = \mathbf{0}$ with optimal parameter vector $w^* = \mathbf{0}$. The $TD(0)$ update equation is then given by

$$w_{t+1} = w_t + \alpha_t(\gamma V(s_{t+1}) - V(s_t)) \ .$$

Then suppose we consider updating the parameter vector using one-step transition samples where the initial state is drawn from the uniform distribution and the next state is given by the transition function. If all weights are initially positive and $\hat{V}(6)$ is initially much larger than the value function estimates at other states, then the value function estimate will begin to diverge. To see this, consider the $TD(0)$ update equation. When $\hat{V}(1), ..., \hat{V}(5)$ are much lower than the value of their successor $\gamma\hat{V}(6)$ and $\hat{V}(6)$ is greater than the value of its successor $\gamma\hat{V}(6)$, then w_0 is increased five times for every one time that it is decreased. Slowly over time, $\hat{V}(6)$ will fall but then quickly become far smaller causing an oscillation in the other direction. This phenomenon is illustrated in the graph below.

Note that in this case the algorithm diverges even though the optimal value function is exactly representable by the given linear function approximation architecture.

Clearly, this example violates two key assumptions in the proof of $TD(0)$'s convergence. Firstly, the matrix Φ is not full rank and thus has infinite possible solutions. Secondly, by sampling according the uniform distribution, the samples

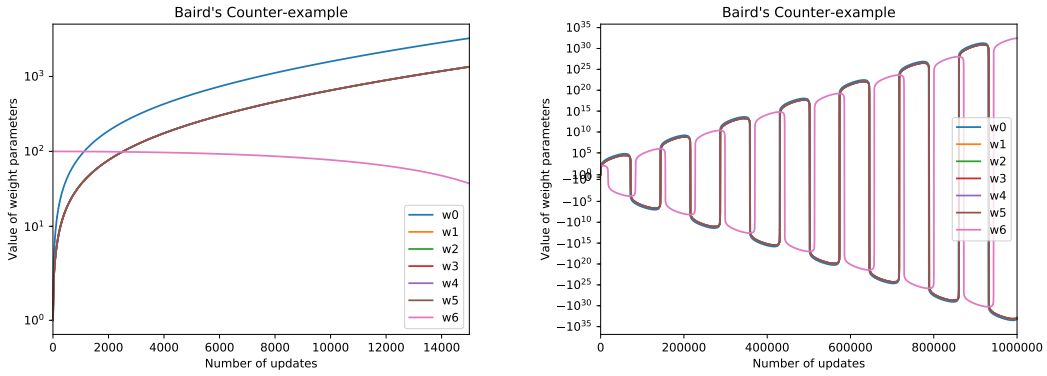

Figure 4.3: Plots of the divergence of parameter values found by $TD(0)$ in Baird's counter-example over 15000 and 1000000 updates and uniform sampling.

are not generated according to the steady state of the state sequence of any policy, which under any policy is to with probability 1 move to state 6.

4.4.2 Tsitsiklis and Van Roy's Counter-example

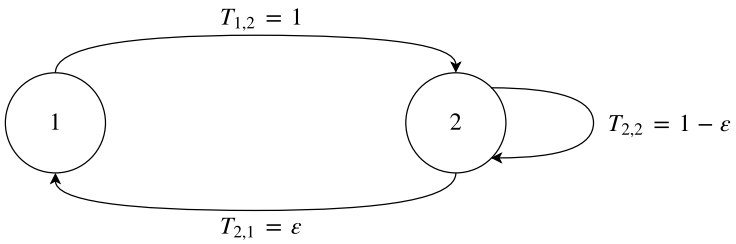

Figure 4.4: Illustration of the MDP underlying Tsitsiklis and Van Roy's counter-example.

Tsitsiklis and Van Roy's counter-example, shown in [3], presents a more analytic perspective as to why $TD(0)$ diverges. Consider the underlying state Markov chain given by the image above where the numbers on the edges represent the probability of the given transition. Note that the transition from state 1 to itself has 0 probability. We again assume that all transitions confer no reward. Thus the exact value function is given by $V^*(\cdot) = 0$. Consider the one-dimension linear approximation architecture given by $\Phi = \begin{bmatrix} 1 & 2 \end{bmatrix}^\top$. Then $\hat{V} = \begin{bmatrix} w & 2w \end{bmatrix}^\top$. Assume that samples are generated uniformly as in Baird's counter-example.

Then computing the expected update gives

$$\begin{aligned}\mathbb{E}[w_{t+1}] &= \mathbb{E}[w_t] + \alpha_t \frac{1}{2}(2\gamma - 1)\mathbb{E}[w_t] + \alpha_t \frac{1}{2} 2\big((1-\epsilon)(2\gamma - 2) + \epsilon(\gamma - 2)\big)\mathbb{E}[w_t] \\ &= \mathbb{E}[w_t] + \frac{\alpha_t}{2}\big(6\gamma - 5 + O(\epsilon)\big)\mathbb{E}[w_t] \ .\end{aligned}$$

Thus for $\gamma > \frac{5}{6}$ and ϵ small, $\mathbb{E}[w_t]$ diverges since then $\frac{\alpha_t}{2}\big(6\gamma - 5 + O(\epsilon)\big) > 1$. In contrast, when $TD(0)$ is updated using samples generated from a trajectory or one-step transition samples from the stationary distribution and transition function, most transitions in state 2 occur from state 2 to itself, giving

$$\mathbb{E}[w_{t+1}] = \mathbb{E}[w_t] + \alpha_t 2(2\gamma - 2)\mathbb{E}[w_t] + O(\epsilon)\mathbb{E}[w_t] \ .$$

In this case, convergence occurs since $\gamma < 1$.

4.5 Summary

In summary, we have seen that the class of natural algorithms is a non-empty and interesting class of algorithms to consider divergence results for. In particular we have seen how three prominent reinforcement learning algorithms - $LSTD$, $TD(0)$, and the residual gradient method - all look to solve for the fixed point of a projected Bellman equation. In the case of $TD(0)$, we have seen examples of divergence. These divergence results show that when one or more of the assumptions set out by Tsitsiklis and Van Roy for the convergence of $TD(0)$ are violated, divergence can occur. More subtly, we have seen some factors that point to the possibility of convergence to the wrong solution. For example, the $TD(0)$ method only minimises the TD error and foregoes the model adequacy component. These factors discussed here provide some of the motivation for the divergence results we seek. As we will see in the next chapter, the dependency on sampling introduces other ways in which solutions may converge to the wrong solution.

Chapter 5

The Ambiguity Conditions

In this chapter we present our main results which we collectively call *the ambiguity conditions*. The ambiguity conditions are a collection of conditions under which projected Bellman equation methods may either diverge or converge to the wrong solution. Throughout the chapter we continue to consider the case where the policy is fixed and so will suppress the superscripts of π. Reward and transition functions R and T will be taken to be defined as $R \coloneqq R^\pi$ and $T \coloneqq T^\pi$. We begin by providing an example to highlight the type of problem captured by our results before going on to show our results in the case of both finite and continuous state spaces.

5.1 A Motivating Example

The following example is drawn from [17]. Consider the two MDPs depicted in Figure 5.1. Unlike most scenarios considered in the rest of this thesis, the policy define here is stochastic rather than deterministic. The edges between states indicate a state transition and the labels indicate the action taken and the reward received. For example, in MDP 1, taking action a_1 in state 1 causes a transition from A to B that incurs 0 reward. When two edges leave a single state, we assume that the policy chooses between actions uniformly. Under such a policy, the transition matrices of MDP 1 and MDP 2 are given by

$$T_1 = \begin{bmatrix} 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}, \quad T_2 = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

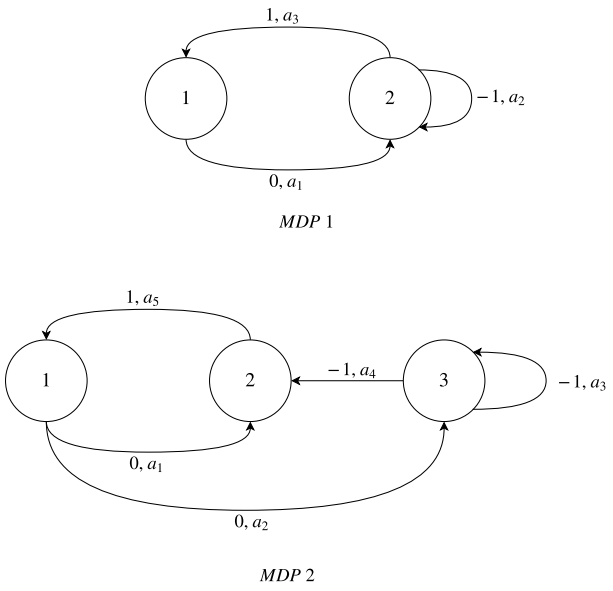

Figure 5.1: Depiction of two MDPs that are observationally equivalent but have different Bellman error and optimal parameter values.

respectively. It can also be seen that the stationary distributions in each MDP are given by $\mu_1 = (\frac{1}{3}, \frac{2}{3})^\top$ and $\mu_2 = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})^\top$ respectively. Now suppose that we use a simple linear function approximation mechanism with a two component parameter vector $w = (w_1, w_2)^\top$. Since MDP 1 has the same number of states as the number of parameters, the value at each state can be represented exactly

$$\hat{V}(1) = w_1 \ , \qquad \hat{V}(2) = w_2 \ .$$

In MDP 2, we assume that states 2 and 3 share a parameter value, i.e.

$$\hat{V}(1) = w_1 \ , \quad \hat{V}(2) = w_2 \ , \quad \hat{V}(3) = w_2 \ .$$

As mentioned in remark 4.5, algorithms that combine sampling techniques with function approximation do not get all the information summarised in the states of the state sequence $\{s_t\}_{t\in\mathbb{N}}$. Instead, the algorithms learn using the feature-vector sequence $\{\phi(s_t)\}_{t\in\mathbb{N}}$ instead.

Now consider the feature vector-reward sequence generated starting in the steady-state. In MDP 1, our function approximation algorithm would see parameter value w_1 followed by 0 with probability $\frac{1}{3}$ and the parameter value w_2 followed by either 1 or -1 with probability $\frac{1}{3}$ each. In MDP 2, we see that the sample distribution is the same. The rewards received in each state is deterministic

and also the steady-state probabilities are equal for all states. Thus, our function approximation algorithm sees w_1 followed by 0 with probability $\frac{1}{3}$ and w_2 followed by 1 or -1 with probability $\frac{1}{3}$ each. Thus, from the perspective of updating via samples, the two MDPs in Figure 5.1 are *indistinguishable*. Furthermore, suppose we use a method that minimises the Bellman error

$$E_{BE} = \left\| (I - \gamma T)\, \Phi\hat{w} - R \right\|_{\mu}^{2} \ .$$

Note that by letting $\hat{V} = \Phi\hat{w}$ the Bellman error can also be expressed as

$$E_{BE} = \sum_{s \in \mathcal{S}} \mu(s) \left[R^{\pi}(s) - \hat{V}(s) + \gamma \mathbb{E}_{\pi} \left[\hat{V}(S_{t+1}) | S_t = s \right] \right]^2 \ .$$

For a parameter value of $\hat{w} = 0$, the Bellman error becomes

$$E_{BE} = \sum_{s \in \mathcal{S}} \mu(s) \left(R^{\pi}(s) \right)^2 \ .$$

Under this parameter value, the Bellman error is 0 in MDP 1 since the expected reward in each state is 0. However, the Bellman error is given by $E_{BE} = \frac{1}{3}(0 + 1 + 1) = \frac{2}{3}$ in MDP 2. Thus the Bellman error is not a unique function of the data sample. This suggests that even though an algorithm minimising the Bellman error may converge, it may converge to the *wrong* parameter vector. Indeed, we note that the optimal parameter vector in MDP 1 is given by $w = 0$. The optimal parameter vector in MDP 2 is however more complex. We note that given our parameter vector, the Bellman error can be expressed as

$$E_{BE} = \frac{1}{3} \left[(-w_1 + \gamma w_2)^2 + (1 - w_2 + \gamma w_1)^2 + \left(-1 + w_2 \left(1 - \gamma \right) \right)^2 \right] \ .$$

We note that since the Bellman error is a sum of quadratic terms, we can characterise its minimiser by settings its gradient to 0

$$\begin{aligned} 0 &= \begin{bmatrix} -(-w_1 + \gamma w_2) + \gamma(1 - w_2 + \gamma w_1) \\ \gamma(-w_1 + \gamma w_2) - (1 - w_2 + \gamma w_1) + (1 - \gamma)(-1 + w_2(1 - \gamma)) \end{bmatrix} \\ &= \begin{bmatrix} w_1(1 + \gamma^2) - 2\gamma w_2 + \gamma \\ -2\gamma w_1 + 2w_2(1 - \gamma + \gamma^2) \end{bmatrix} \ . \end{aligned}$$

Clearly, the solution that minimises the Bellman error is a complicated function of γ. However taking the limit as $\gamma \to 1$, we obtain two equivalent equations

$$\begin{aligned} 2w_1 - 2w_2 + 1 &= 0 \\ -2w_1 + 2w_2 - 1 &= 0 \ . \end{aligned}$$

Since there are two variables and only one equation, there exists infinitely many parameter vectors that satisfy this equation. Now suppose that MDP 1 was the true environment. Then any sample-based, function approximation method looking to minimise the Bellman error would not be able to distinguish whether the underlying environment was truly MDP 1 or MDP 2. Thus, it may converge to any one of the parameter vectors that satisfy MDP 2's Bellman error constraints instead. In this case, the approximate value function $\hat{V} = \Phi w$ will be different from the true value function of $\hat{V} = 0$.

The scenario here presents a subtle form of divergence that can occur for when dealing with function approximation methods. Instead of the algorithms explicitly diverging, we have seen how it is possible that they may converge but to the *wrong* solution. Our main results, aptly named the ambiguity conditions, in the next section look to generalise this phenomenon to all natural algorithms and provide sufficient conditions that capture exactly when this phenomenon may occur. Recall that a key property of natural algorithms is that their solution is characterised as the fixed point of a projected Bellman equation. Furthermore, the solution to a projected Bellman equation is completely determined by the three quantities A, B, and b (see corollaries 3.5 and 4.2). Our ambiguity conditions show that there exist scenarios where different environments with different parameter vectors 'look' the same under projection, that is the quantities A, B, and b are the same for the different environments, but the true parameter vector is different.

5.2 The Ambiguity Conditions in Finite State Space

In this section we look to derive a series of ambiguity conditions in the case of a finite state space. Throughout we will assume that the state space is given by $\mathcal{S} = \{1, \ldots, n\}$, where n is 'large', and the action space $\mathcal{A}$ is finite. Also, we assume that our other key assumptions hold. We assume that we are looking for an approximate solution in a subspace with dimension $k < n$, and that assumption 5 holds, that is the state Markov chain admits a stationary distribution μ such that $\mu(s) > 0$ for all $s \in \mathcal{S}$. Each condition is defined in relation to the previous and was done so in order to try find a constructive result. As a result, we are able to construct an example where ambiguity holds. To begin with, we first define the *Bellman template* which will form the framework of our analysis.

Definition 5.1 (Bellman template)**.** Suppose Π_Ψ, a natural projection operator characterised by (Φ, Ψ). Let $w \in \mathbb{R}^k$, $R : \mathcal{S} \to \mathbb{R}$, and $T : \mathcal{S} \times \mathcal{S} \to [0, 1]$ be variables. We define the following constraints collectively as the *Bellman template*:

- (w, R, T) satisfy the Bellman equations

$$\Phi w = R + \gamma T \Phi w \ . \tag{5.1}$$

- Let $A \in \mathbb{R}^{k\times k}$, $B \in \mathbb{R}^{k\times k}$, and $b \in \mathbb{R}^k$ be given by

$$A = \Psi^\top D_\mu \Phi \ , B \qquad\qquad = \Psi^\top D_\mu T \Phi \ , b = \Psi^\top D_\mu R \ .$$

  respectively.

We say that a tuple (w^*, R^*, T^*) is a solution to the Bellman template if it satisfies these constraints.

Remark 5.2. Recall that a solution to a projected Bellman equation is uniquely determined by A, B, b (see corollary 4.2). Thus it may seem strange that we could have multiple solutions. The crucial difference however is that we now let R and T vary. The first constraint can be seen as requiring that the value function for the fixed policy be *exactly* representable in our approximation subspace. The second constraint provides a condition to consider when different tuples (w, R, T) produce the same quantities A, B, b.

Our next assumption asserts a true environment and parameter vector to which other incorrect solutions can be compared to.

Assumption 4. Let $w^* \in \mathbb{R}^k$ be a parameter vector, and $R^* : \mathcal{S} \to \mathbb{R}$ and $T^* : \mathcal{S} \times \mathcal{S} \to [0, 1]$ be an immediate reward function and transition function respectively such that (w^*, R^*, T^*) is a solution to the Bellman template. We assume that (w^*, R^*, T^*) is the true solution to the Bellman template.

Under assumption 4, if the Bellman template has at least one other solution with a different parameter vector, all natural algorithms characterised by the given projection Π_Ψ may either diverge or converge to the wrong parameter vector. We thus define ambiguity as follows.

Definition 5.3 (Ambiguity)**.** If the Bellman template has more than one solution and at least two of the solutions have different parameter vectors, then we say that ambiguity holds.

Having defined ambiguity, we can now proceed to show our first ambiguity condition in the case of a finite state space.

Theorem 5.4. *Suppose assumption 4 holds. Then ambiguity holds if and only if there exists* $v \neq 0 \in \mathbb{R}^k$ *such that* $(A - \gamma B)\, v = 0$.

Proof. Suppose ambiguity holds, that is there exists another solution $(w^\circ, R^\circ, T^\circ)$ satisfying the Bellman template such that $w^\circ \neq w^*$. Then $(w^\circ, R^\circ, T^\circ)$ satisfies

$$(A - \gamma B)\, w^\circ = b \ .$$

To see, this note that since $(w^\circ, R^\circ, T^\circ)$ is a solution to the Bellman template, T° and R° satisfy A, B and b. Since $(w^\circ, R^\circ, T^\circ)$ also satisfies equation (5.1), we have by multiplying equation (5.1) on the left by $\Psi^\top D_\mu$

$$\Psi^\top D_\mu \Phi w^\circ = \Psi^\top D_\mu R + \gamma \Psi^\top D_\mu T^\circ \Phi w^\circ \ .$$

Re-arranging and substituting in the definitions of A, B, and b gives $(A - \gamma B)\, w^\circ = b$. Then similarly, $(A - \gamma B)\, w^* = b$. Letting $v = w^1 - w^*$, we have

$$(A - \gamma B)\left(w^1 - w^*\right) = b - b \ ,$$

which gives

$$(A - \gamma B)\, v = 0 \ .$$

We now show the reverse. Suppose that A, B, and b are given as in the Bellman template and that there exists $v \neq 0 \in \mathbb{R}^k$ such that $(A - \gamma B)\, v = 0$. For any $\lambda > 0$, let

$$\begin{aligned} w^\lambda &\coloneqq w^* + \lambda v \ , \\ T^\lambda &= T^* \ , \\ R^\lambda &\coloneqq \left(I - \gamma T^\lambda\right) \Phi w^\lambda \ . \end{aligned}$$

We now show that $(w^\lambda, R^\lambda, T^\lambda)$ satisfies the Bellman template. Firstly, it satisfies equation (5.1) since the left-hand side is given by

$$\Phi w^\lambda = \Phi w^* + \lambda \Phi v \ ,$$

and the right-hand side is given by

$$\begin{aligned} R^\lambda + \gamma T^\lambda \Phi w^\lambda &= \left(I - \gamma T^*\right) \Phi \left(w^* + \lambda v\right) + \gamma T^\lambda \Phi \left(w^* + \lambda v\right) \\ &= \Phi w^* + \lambda \Phi v \ . \end{aligned}$$

Finally, let $A^\lambda = \Psi^\top D_\mu \Phi$, $B^\lambda = \Psi^\top D_\mu T^\lambda \Phi$ and $b^\lambda = \Psi^\top D_\mu R^\lambda$. Clearly $A^\lambda = A$ and $B^\lambda = B$. From equation (5.1) we have

$$\left(I - \gamma T^\lambda\right) \Phi w^\lambda = R^\lambda \ .$$

Multiplying this equation $\Psi^\top D_\mu$ from the left gives

$$\begin{aligned} b^\lambda &= \Psi^\top D_\mu \left(I - \gamma T^\lambda\right) \Phi w^\lambda \\ &\overset{(a)}{=} (A - \gamma B)\, w^\lambda \\ &= (A - \gamma B)\, (w^* + \lambda v) \\ &\overset{(b)}{=} (A - \gamma B)\, w^* \\ &= b \ , \end{aligned}$$

where (a) follows by definition of A and B since $T^\lambda = T^*$ and (b) follows by our starting assumption. Thus, all constraints in the Bellman template are satisfied and $(w^\lambda, R^\lambda, T^\lambda)$ is a solution to the Bellman template for all $\lambda > 0$. Thus, ambiguity holds. □

Remark 5.5. From our proof of theorem 5.4, if $(A - \gamma B)\, v = 0$ holds, there exists an infinite number of possible solutions to the projected Bellman equation given by the projection Π_Ψ. In particular we have an explicit construction for a set of reward functions and parameter vectors that satisfy the Bellman template. We now look to derive equivalent conditions that give us a construction for T.

The following corollary provides a useful re-definition that will help us in finding conditions to construct T.

Corollary 5.6. *Suppose assumption 1 holds. Let $\chi_i(s) \coloneqq \psi_i(s)\mu(s)$ for all $s = 1, \ldots, n$, where ψ_i is the i^{th} column of Ψ, and let $\varphi = \Phi v$ for $v \neq 0 \in \mathbb{R}^k$. Then ambiguity holds if and only if there exists T such that for all i,*

$$\chi_i^\top \left(I - \gamma T\right) \varphi = 0 \ . \tag{5.2}$$

Proof. Suppose T exists such that $\Psi^\top D_\mu T \Phi = B$. Expanding $(A - \gamma B)v = 0$ gives

$$\Psi^\top D_\mu \left(I - \gamma T\right) \Phi v = 0 \ .$$

Note that $\left[\Psi^\top D_\mu\right]_{i,s} = \psi_i(s)\mu(s) = \chi_i(s)$. Thus for all i,

$$\chi_i^\top \left(I - \gamma T\right) \Phi v = 0 \ .$$

Substituting in $\varphi = \Phi v$ gives $\chi_i^\top \left(I - \gamma T\right) \varphi = 0$. □

Remark 5.7. Since v can be interpreted as the error in the parameter vector, the vector φ can be treated as the error in the value function between different the value function of the true solution and that of an erroneous solution that satisfies the Bellman template.

Ambiguity under an orthogonal basis

In this section we look to derive ambiguity conditions in the case where the set $\psi_1, \ldots, \psi_k$, the columns of Ψ that characterise the space that a natural projection projects orthogonally to, are orthogonal to each other (with respect to $\|\cdot\|_\mu$) and are all non-negative vectors. This is formalised in the following assumption.

Assumption 5. Assume that for all $i = 1, \ldots, k$ and $s = 1, \ldots, n$ that $\psi_i(s) \geq 0$. Furthermore, assume that $\psi_1, \ldots, \psi_k$ forms an orthonormal basis (with respect to $\|\cdot\|_\mu$), i.e. $\langle \psi_i, \psi_j \rangle_\mu = 0$ for all $i \neq j$ and $\|\psi_i\|_\mu = 1$ for all i.

Remark 5.8. As a direct consequence of assumption 5, if $\psi_i(s) \neq 0$ for a given s, then $\psi_j(s) = 0$ for all $j \neq i$.

The next theorem helps us toward the next theorem which gives us an explicit construction for T.

Theorem 5.9. *Suppose assumptions 4 and 5 hold. Let $t_i \in \mathbb{R}^n$ be vector such that $\sum_{s=1}^n t_i(s) = 1$ and $t_i(s) \geq 0$ for all $s = 1, \ldots, n$. Then ambiguity holds if and only if there exists $\varphi \neq 0 \in span(\Phi)$ such that for all $i = 1, \ldots, k$, there exists t_i such that*

$$\chi_i^\top \varphi = \gamma t_i^\top \varphi \ .$$

Proof. We first note that ambiguity holds if and only if equation there exists T such that 5.2 holds. Then re-arranging equation 5.2 gives

$$\chi_i^\top \varphi = \gamma \chi_i^\top T \varphi \ .$$

Since $\|\psi_i\|_\mu = 1$ for all i, $\|\chi_i\| = 1$. Then we can simply take $t_i = \chi_i^\top T$. □

We are now able to present the main result under assumption 4.

Theorem 5.10. *Assume that assumptions 4 and 5 hold. Then ambiguity holds if and only if there exists $\varphi \neq 0 \in span(\Phi)$ such that for all i,*

$$\gamma \varphi_{\min} \leq \chi_i^\top \varphi \leq \gamma \varphi_{\max} \ .$$

Proof.
$\Longrightarrow$:
Assume that ambiguity holds. Then by theorem 5.9, $\chi_i^\top \varphi = \gamma t_i^\top \varphi$ for some t_i. Define T as

$$T_{s,s'} = t_i(s') \text{ if } \chi_i(s) \neq 0 \ .$$

For a given $s \in \{1, \ldots, n\}$, under assumption 5 and by the definition of χ_i, there is a unique i where $\chi_i(s) \neq 0$. The definition of T allows any t_i to be satisfied since $\chi_i^\top T = t_i^\top$.

Let us define the following:

$$\varphi_{\max} = \sup_s \varphi(s) \quad \text{and} \quad \varphi_{\min} = \inf_s \varphi(s)$$

Since s takes values in $\{1, \ldots n\}$, the supremum and infimum are just the maximum and minimum and are guaranteed to exist. We also define $s_{\max} = \arg\max_s \varphi(s)$ and $s_{\min} = \arg\min_s \varphi(s)$. Then $t_i(s') = \mathbf{1}(s_{\max} = s')$ and $t_i(s') = \mathbf{1}(s_{\min} = s')$ achieve the maximum and minimum values for $t_i^\top \varphi$ as:

$$\sup_{t_i} t_i^\top \varphi = \sum_{s'} \mathbf{1}(s_{\max} = s')\varphi(s') = \varphi_{\max}$$
$$\inf_{t_i} t_i^\top \varphi = \sum_{s'} \mathbf{1}(s_{\min} = s')\varphi(s') = \varphi_{\min}$$

Since $\gamma \inf_{t_i} t_i^\top \varphi \leq \gamma t_i^\top \varphi \leq \gamma \sup_{t_i} t_i^\top \varphi$ we have that $\gamma\varphi_{\min} \leq t_i^\top \varphi \leq \gamma\varphi_{\max}$.

$\Longleftarrow$:
Assume for any i that $\varphi_{\min} \leq \frac{1}{\gamma}\chi_i^\top \varphi \leq \varphi_{\max}$. We note that $\chi_i^\top \varphi(s)$ is a continuous function. Thus, we can define t_i as

$$t_i(s') = \frac{\frac{1}{\gamma}\chi_i^\top \varphi - \varphi_{\min}}{\varphi_{\max} - \varphi_{\min}} \delta_{s_{\max}}(s') + \frac{\varphi_{\max} - \frac{1}{\gamma}\chi_i^\top \varphi}{\varphi_{\max} - \varphi_{\min}} \delta_{s_{\min}}(s')$$

Thus,

$$\begin{aligned} t_i^\top \varphi &= \frac{\frac{1}{\gamma}\chi_i^\top \varphi - \varphi_{\min}}{\varphi_{\max} - \varphi_{\min}} \varphi_{\max} + \frac{\varphi_{\max} - \frac{1}{\gamma}\chi_i^\top \varphi}{\varphi_{\max} - \varphi_{\min}} \varphi_{\min} \\ &= \frac{1}{\gamma}\chi_i^\top \varphi \ . \end{aligned}$$

Thus, $\chi_i^\top \varphi = \gamma t_i^\top \varphi$. $\square$

Remark 5.11. From the proof of 5.10, we are able to construct a T using t_i. Also, we are able to relate γ to the function φ.

Example 5.12. In this example we give an explicit construction of multiple environments that satisfy the Bellman template under assumption 5. Consider the scenario where $\mathcal{S} = \{1, 2\}$ and the stationary distribution is given by $\mu = (\frac{1}{2}, \frac{1}{2})^\top$. Since we consider lower-dimensional approximations, suppose that Φ and Ψ are defined as follows

$$\Phi = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \ , \quad \Psi = \begin{bmatrix} \frac{2}{3} \\ \frac{4}{3} \end{bmatrix} \ .$$

Let $\varphi = \Phi$. Given $\Psi = [\psi_1]$, χ_1 is given by $\chi_1 = (\frac{1}{3}, \frac{2}{3})^\top$. Given φ, $\varphi_{\max} = 2$ and $\varphi_{\min} = 1$. Also $\chi_1^\top \varphi = \frac{5}{3}$. Thus, under theorem 5.10, ambiguity holds if and only if

$$\gamma \leq \frac{5}{3} \leq 2\gamma \ .$$

For $\gamma \in [\frac{5}{6}, 1]$, this condition holds. For $\gamma = \frac{5}{6}$, we find that there exists t_1 that satisfies theorem 5.9, namely $t_1 = [0 \ 1]^\top$. Thus, as done in the proof of 5.10, we are able to define our transition matrix according to

$$T_{s,s'} = t_i(s') \ \text{ if } \chi_i(s) \neq 0 \ .$$

Thus, we define T as

$$T = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \ .$$

Given we chose $\varphi = \Phi$, we let $v = 1$. Then we choose simply as $w^\lambda = \lambda$. Now from the proof of theorem 5.4, we define R^λ by

$$\begin{aligned} R^\lambda &= (I - \gamma T)\, \Phi w^\lambda \\ &= \begin{bmatrix} 1 & -\frac{5}{6} \\ 0 & \frac{1}{6} \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} \lambda \\ &= \frac{\lambda}{6} \begin{bmatrix} -2 \\ 1 \end{bmatrix} \ . \end{aligned}$$

By construction, $(w^\lambda, R^\lambda, T)$ satisfy the Bellman template for all values of $\lambda > 0$. Thus, any natural algorithm characterised by (Φ, Ψ) as defined in our example may either diverge or converge to the wrong solution. Indeed, suppose

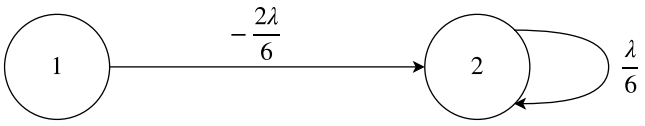

Figure 5.2: An MDP where any value of $\lambda > 0$ characterises a different environment. With $w^\lambda = \lambda$, all these environments look the same to natural algorithms whose solutions are characterised by $\Phi = [1\ 2]^\top$ and $\Psi = [\frac{2}{3}\ \frac{4}{3}]^\top$.

that the true solution is given when $\lambda = 0$. Then the true value function is given by $V^0 = 0$. Suppose that V^λ is the value function for w^λ. Then as λ approaches infinity, $\left|V^\lambda - V^0\right| \to \infty$. Thus, the approximation found by any natural algorithm characterised by (Φ, Ψ) can be arbitrarily bad.

Summary

Our results in this section show sufficient conditions under which ambiguity holds. As shown in example 5.12, our results allow for the construction of environments that all satisfy the Bellman template. These results so far however are limited to the case where the state space is finite. Furthermore, theorem 5.10 only holds under assumption 5. Further results for the finite state space were not found. In the next section we turn return to the case of a continuous state space and show that in this case, analogous conditions were found and assumption 5 can eventually be lifted.

5.3 The Ambiguity Conditions in Continuous State Space

In this section, we look to show conditions under which ambiguity holds in the case of a continuous state space. We return to the set-up described in chapters 2 and 3. In particular, the state space $\mathcal{S}$ is a compact subspace of $\mathbb{R}$ and the action space $\mathcal{A}$ is kept finite. We will continue to assume that the state Markov process is ergodic and admits a stationary distribution μ such that $\mu(s) > 0$ for all $s \in \mathcal{S}$. Under these conditions, our set of natural projection operators are defined as in 3.3 and our set of natural algorithms is defined as in definition 3.6. Many of the

results that hold in the case of a finite state space hold analogously in this case. Furthermore, we are able to show a condition under which *any* natural algorithm will face ambiguity. Due to the significant added complexity when a continuous state space is introduced, no explicit examples were found. In a similar fashion to the case presented for the finite state space, we will define the Bellman template which forms the basis of our analysis. Ambiguity is then defined as in definition 5.3. Thus in this case, natural algorithms will again either diverge or converge to the wrong solution under ambiguity.
We now define the Bellman template.

Definition 5.13 (Bellman template)**.** Suppose Π_Ψ, a natural projection operator characterised by (Φ, Ψ) is given. Let $w \in \mathbb{R}^k$, $R : \mathcal{S} \to \mathbb{R}$, and $T : \mathcal{S} \times \mathcal{S} \to [0, 1]$ be variables. We define the following constraints collectively as the *Bellman template*:

- (w, R, T) satisfy the Bellman equations

$$\phi^\top w = R + \gamma P_T \phi^\top w \ . \tag{5.3}$$

- Let $A \in \mathbb{R}^{k\times k}$, $B \in \mathbb{R}^{k\times k}$, and $b \in \mathbb{R}^k$ be given by

$$\begin{aligned} A_{ij} &= \langle \psi_i, \phi_j \rangle_\mu \ , \quad i, j = 1, \ldots, k \\ B_{ij} &= \langle \psi_i, P_T \phi_j \rangle_\mu \ , \quad i, j = 1, \ldots, k \\ b_i &= \langle \psi_i, R \rangle_\mu \ . \end{aligned}$$

  respectively.

We say that a triple $(w^\circ, R^\circ, T^\circ)$ is a solution to the Bellman template if it satisfies these constraints.

Assumption 6. Let $w^* \in \mathbb{R}^k$ be a parameter vector, and $R^* : \mathcal{S} \to \mathbb{R}$ and $T^* : \mathcal{S} \times \mathcal{S} \to [0, 1]$ be an immediate reward function and transition function respectively. We assume that (w^*, R^*, T^*) is a solution to the Bellman template.

We begin by deriving analogous results to those found in the case of a finite state space.

Theorem 5.14. *Suppose that assumption 6 holds. Then ambiguity holds if and only if there exists $v \neq 0 \in \mathbb{R}^k$ such that*

$$(A - \gamma B)v = 0 \ .$$

Proof. Suppose ambiguity holds. Then there exists another solution to the Bellman template which we will denote by $(w^\circ, R^\circ, T^\circ)$ such that $w^\circ \neq w^*$. By corollary 3.5, we have that both $w = w^\circ$ and $w = w^*$ satisfy

$$(A - \gamma B)w = 0 \ .$$

Then trivially, $(A-\gamma B)v = 0$ holds since we can take $v = w^* - w^\circ$. Now consider the reverse. Assume that $(A - \gamma B)v = 0$ holds for some $v \neq 0$. For $\lambda > 0$, let us define

$$\begin{aligned} w^\lambda &= w^* + \lambda v \ , \\ T^\lambda &= T^* \ , \\ R^\lambda(\cdot) &= \phi^\top(\cdot)w^\lambda - \gamma P_{T^\lambda}\phi^\top(\cdot)w^\lambda \ . \end{aligned}$$

Then $(w^\lambda, R^\lambda, T^\lambda)$ satisfies the Bellman equation since

$$\begin{aligned} \mathcal{T}\phi^\top(\cdot)w &= R^\lambda(\cdot) + \gamma P_{T^\lambda}\phi^\top(\cdot)w^\lambda \\ &= \phi^\top(\cdot)w^\lambda \ , \end{aligned}$$

which is precisely the left-hand side of the Bellman equation. Let $A^\lambda \in \mathbb{R}^{k\times k}$, $B^\lambda \in \mathbb{R}^{k\times k}$ and $b \in \mathbb{R}^k$ be given by

$$\begin{aligned} A^\lambda_{ij} &= \langle \psi_i, \phi_j \rangle_\mu \ , \quad i,j = 1,2,\ldots,k \ , \\ B^\lambda_{ij} &= \langle \psi_i, P_{T^\lambda}\phi_j \rangle_\mu \ , \quad i,j = 1,2,\ldots,k \ , \\ b^\lambda_i &= \langle \psi_i, R \rangle_\mu \ , \quad i,j = 1,2,\ldots,k \ . \end{aligned}$$

We trivially have that $A^\lambda = A$ and we note that $B^\lambda = B$ since $T^\lambda = T^*$. By corollary 3.5, applying the projection Π_Ψ characterised by (Φ, Ψ) onto the Bellman equations induces the following

$$\left(A^\lambda - \gamma B^\lambda\right) w^\lambda = b^\lambda \ .$$

We then have the following derivation

$$\begin{aligned} b^\lambda &= \left(A^\lambda - \gamma B^\lambda\right) w^\lambda \\ &\overset{(a)}{=} \left(A^\lambda - \gamma B^\lambda\right) (w^* + \lambda v) \\ &\overset{(b)}{=} \left(A^\lambda - \gamma B^\lambda\right) w^* \\ &= b \ . \end{aligned}$$

Thus we have that $b^\lambda = b$ as well. Thus, for any $\lambda > 0$, $(w^\lambda, R^\lambda, T^\lambda)$ satisfies the Bellman equation as well and so ambiguity holds. □

Thus, the result holds analogously to theorem 5.4 in the finite case. In particular, under theorem 5.14 there are infinitely many solutions to the Bellman template. We now re-define a few terms and show a corollary result.

Corollary 5.15. *Assume that assumption 6 holds. For all $i = 1, \dots, k$, let $\chi_i \coloneqq \psi_i(\cdot)\mu(\cdot)$ where μ is the stationary distribution, and let $\varphi(\cdot) \coloneqq \phi^\top(\cdot)v$ for $v \neq 0 \in \mathbb{R}^k$. Then ambiguity holds if and only if there exists $\varphi \neq 0 \in im(\Phi)$ and T such that for all $i = 1, \dots, k$,*

$$\int_{\mathcal{S}} \chi_i(s)\big(I - \gamma P_T\big)\varphi(s)ds = 0 \ .$$

Proof. From theorem 5.14, we have that ambiguity holds if and only if there exists $v \neq 0 \in \mathbb{R}^k$ such that

$$(A - \gamma B)\, v = 0 \ .$$

Now for all $i, j = 1, \dots, k$

$$\begin{aligned} A_{ij} - \gamma B_{ij} &= \int_{\mathcal{S}} \psi_i(s)\mu(s)\phi_j(s) - \gamma\psi_i(s)\mu(s)P_T\phi_j(s)ds \\ &= \int_{\mathcal{S}} \chi_i(s)\,(I - \gamma P_T)\,\phi_j(s)ds \ . \end{aligned}$$

Now note that $A_{i\cdot}$ and $B_{i\cdot}$ are row vectors of A and B for all $i = 1, \dots, k$. Then since $(A - \gamma B)\, v = 0$, we have for all $i = 1, \dots, k$ the following derivation

$$\begin{aligned} 0 &= (A_{i\cdot} - \gamma B_{i\cdot})\, v \\ &= \sum_{j=0}^{k} \big(A_{ij} - \gamma B_{ij}\big)\, v_j \\ &\overset{(a)}{=} \sum_{j=0}^{k} \int_{\mathcal{S}} \chi_i(s)\,(I - \gamma P_T)\,\phi_j(s)v_j ds \\ &\overset{(b)}{=} \int_{\mathcal{S}} \chi_i(s)\,(I - \gamma P_T) \sum_{j=0}^{k} \phi_j(s)v_j ds \\ &= \int_{\mathcal{S}} \chi_i(s)\,(I - \gamma P_T)\,\varphi(s)ds \ . \end{aligned}$$

Here in (a) we substituted in the derivation of $A_{ij} - \gamma B_{ij}$ from above and in (b) we used the Tonelli-Fubini theorem to swap the sum and the integral. Since we've only looked at equivalences, our if and only if result holds. □

5.3.1 Ambiguity for Orthogonal Ψ

In this section we place some restriction on the set Ψ, namely that it is an orthonormal set of positive basis functions that span $im(\Pi^*_\Psi)$. We formalise these restrictions in the next assumption.

Assumption 7. Assume that $\Psi \coloneqq \{\psi_1, \ldots, \psi_k\}$ is an orthonormal basis for $im(\Pi^*_\Psi)$. Also assume that for all $i = 1, \ldots, k$, $\psi_i \geq 0$ and $\|\psi_i\|_{1,\mu} \coloneqq \int_{\mathcal{S}} |\psi_i| \, \mu(s) = 1$.

Remark 5.16. Note that the assumption that $\|\psi_i\|_{1,\mu} = 1$ is not restrictive. This is since $\psi_i \in im(\Pi^*_\Psi) \subset L^2(\mathcal{S}, \mu)$ and so $\|\psi_i\|_\mu < \infty$ implying that $\|\psi_i\|_{1,\mu}$ is also bounded. Thus we can normalise ψ_i such that $\int_{\mathcal{S}} \psi_i \mu(s) = 1$ holds. Thus, $\int_{\mathcal{S}} \chi_i(s) = \int_{\mathcal{S}} \psi_i(s)\mu(s) = 1$. Assumption 7 also implies that for any $s \in \mathcal{S}$, $\psi_i(s) \neq 0$ implies $\psi_j(s) = 0$ for all $j \neq i$. To see this note that since $\mu(s) > 0$ for all $s \in \mathcal{S}$, $\langle \psi_i, \psi_j \rangle_\mu = 0$ must mean that either $\psi_i = 0$ or $\psi_j = 0$ for all $i \neq j$.

We now have the following result.

Theorem 5.17. *Assume that assumptions 6 and 7 hold. Then ambiguity holds if and only if there exists $\varphi \neq 0 \in im(\Pi_\Psi)$ where for all i, there exists a a function $t_i : \mathcal{S} \to [0, 1]$, $\int_{\mathcal{S}} t_i(s)ds = 1$ such that*

$$\int_{\mathcal{S}} \chi_i(s)\varphi(s)ds = \gamma \int_{\mathcal{S}} t_i \varphi(s)ds \ .$$

Proof. Note that for all $i = 1, \ldots k$, corollary 5.15 gives

$$\int_{\mathcal{S}} \chi_i(s)\big(I - \gamma P_T\big)\varphi(s)ds = 0 \ ,$$

which after re-arranging gives

$$\int_{\mathcal{S}} \chi_i(s)\varphi(s)ds = \gamma \int_{\mathcal{S}} \chi_i(s) P_T \varphi(s)ds \ .$$

We have after expanding the right-hand side

$$\gamma \int_{\mathcal{S}} \chi_i(s) P_T \varphi(s)ds = \gamma \int_{\mathcal{S}} \chi_i(s) \int_{\mathcal{S}} T(s'|s)\varphi(s')ds'ds \ .$$

Under assumption 7, for each $s \in \mathcal{S}$ there is a unique $i = 1, \ldots, k$ such that $\chi_i(s) \neq 0$ since $\chi_i(s) \neq 0$ implies that $\chi_j(s) = 0$ for all $j \neq i$. Let t_i be defined

such that $T(s'|s) = t_i(s')$ for the unique i where $\chi_i(s) \neq 0$. Then the right-hand side becomes

$$\gamma \int_{\mathcal{S}} \chi_i(s) \int_{\mathcal{S}'} T(s'|s)\varphi(s')ds'ds = \gamma \int_{\mathcal{S}} \chi_i(s) \int_{\mathcal{S}'} t_i(s')\varphi(s')ds'ds$$
$$\overset{(a)}{=} \gamma \int_{\mathcal{S}'} t_i(s')\varphi(s')ds' \ .$$

Here in (a) we again use the Tonelli-Fubini theorem to swap the integrals as well as the fact that $\int_{\mathcal{S}} \chi_i(s)ds = \int_{\mathcal{S}} \psi_i(s)\mu(s)ds = 1$. □

Before deriving our next result, recall that we have assumed $\mathcal{S}$ to be compact. Thus continuous functions over $\mathcal{S}$ achieve a maximum and minimum value over $\mathcal{S}$.

Theorem 5.18. *Assume that assumptions 6 and 7 hold. Then ambiguity holds if and only if there exists $\varphi \neq 0 \in im(\Pi_\Psi)$, and T such that for all $i = 1, \dots k$,*

$$\gamma\varphi_{\min} \leq \int_{\mathcal{S}} \chi_i(s)\varphi(s)ds \leq \gamma\varphi_{\max}$$

where $\varphi_{\max} = \max_s \varphi(s)$ and $\varphi_{\min} = \min_s \varphi(s)$.

Proof. Suppose that ambiguity holds. Then by theorem 5.17, there exists $\varphi \neq 0$ where for all $i = 1, \dots k$ there exists a $t_i : \mathcal{S} \to [0, 1]$, $\int_{\mathcal{S}} t_i(s)ds = 1$ such that

$$\int_{\mathcal{S}} \chi_i(s)\varphi(s)ds = \gamma \int_{\mathcal{S}} t_i(s)\varphi(s)ds \ .$$

Now since $\int_{\mathcal{S}} t_i(s)ds = 1$, $\varphi_{\min} \leq \int_{\mathcal{S}} t_i(s)\varphi(s)ds \leq \varphi_{\max}$. Thus

$$\gamma\varphi_{\min} \leq \gamma \int_{\mathcal{S}} t_i(s)\varphi(s)ds \leq \gamma\varphi_{\max}$$

and substituting in $\int_{\mathcal{S}} \chi_i(s)\varphi(s)ds$ gives the desired inequality. Now consider the case where $\varphi_{\min} \leq \frac{1}{\gamma} \int_{s\in S} \chi_i(s)\varphi(s)ds \leq \varphi_{\max}$. Again, since we require $\int_{\mathcal{S}} t_i(s)ds = 1$, $\int_{\mathcal{S}} t_i(s)\varphi(s)ds$ is bounded between $\varphi_{\min}$ and $\varphi_{\max}$. Thus

$$\gamma\varphi_{\min} \leq \gamma \int_{\mathcal{S}} t_i(s)\varphi(s)ds \leq \gamma\varphi_{\max} \ .$$

Thus since we can pick t_i however we like, by continuity we can choose t_i such that $\int_{\mathcal{S}} \chi_i(s)\varphi(s)ds = \gamma \int_{\mathcal{S}} t_i(s)\varphi(s)ds$. Thus the result holds. □

Thus, up to this point we have analogous results that hold when the state space is either finite or continuous. In the next section we derive our most general condition for ambiguity.

5.3.2 Ambiguity for Increasing Discount Factor

As we saw in the motivating example, minimising the Bellman error as γ approached 1 caused their to be an infinite number of solutions. We are thus motivated to consider this particular set-up. We first define a key property of the value error function.

Definition 5.19.
Let φ be a function in $im(\Pi_\Psi)$ such that $\varphi \neq 0$. Let $\mathcal{N}_\gamma \coloneqq \{s : \varphi(s) \geq \gamma\varphi_{\max}\}$. Then we say that φ has *non-flat maximum* if $\mu[\mathcal{N}_\gamma] = 0$. Similarly, define $\mathcal{N}_\gamma^- \coloneqq \{s : \varphi(s) \leq \gamma\varphi_{\min}\}$. We say that φ has *non-flat minimum* if $\mu[\mathcal{N}_\gamma^-] = 0$.

The following lemma defines an equivalent condition.

Lemma 5.20. *Assume that assumptions 6 and 7 hold. Let φ be a function in $im(\Pi_\Psi)$ such that $\varphi \neq 0$. Then φ has non-flat maximum if and only if $\mu[\mathcal{N}_\gamma] \to 0$ as $\gamma \to 1$. Similarly, φ has non-flat minimum if and only if $\mu[\mathcal{N}_\gamma^-] \to 0$ as $\gamma \to 1$.*

Proof. We will only prove that φ has non-flat maximum if and only if $\mu[\mathcal{N}_\gamma] \to 0$ as $\gamma \to 1$ noting that adapting the proof for the non-flat minimum case is routine.

Suppose $\mu[\mathcal{N}_1] = 1$. Then we have that

$$\begin{aligned} 1 &= \mu[\mathcal{S}\backslash\mathcal{N}_\gamma + \mathcal{N}_\gamma] \\ &= \mu[\mathcal{S}\backslash\mathcal{N}_\gamma] + \mu[\mathcal{N}_\gamma] \ . \end{aligned}$$

Re-arranging and taking the limit as γ goes to 1 gives

$$\begin{aligned} \lim_{\gamma\to 1} \mu[\mathcal{N}_\gamma] &= \lim_{\gamma\to 1} 1 - \mu[S\backslash\mathcal{N}_\gamma] \\ &= 0 \end{aligned}$$

and the result holds. Now suppose $\mu[\mathcal{N}_\gamma] \to 0$ as $\gamma \to 1$. Then trivially $\lim_{\gamma\to 1} \mu[\mathcal{N}_\gamma] = \mu[\mathcal{N}_1] = 0$. Thus the result holds. □

The next theorem is the main result in the case where we have an orthonormal set of positive basis functions.

Theorem 5.21. *Assume that assumptions 6 and 7 hold. Let $\varphi \neq 0$ be an element of $im(\Pi_\Psi)$. Then if φ has non-flat maximum and minimum, ambiguity holds.*

Proof. Suppose φ has a non-flat maximum. If $\varphi_{\max} \leq 0$, then $\gamma\varphi_{\max} \geq \varphi_{\max}$ for all γ, and thus

$$\begin{aligned}\int_{\mathcal{S}} \chi_i(s)\varphi(s)ds &\leq \varphi_{\max} \int_{\mathcal{S}} \chi_i(s)ds \\ &\overset{(a)}{=} \varphi_{\max} \\ &\leq \gamma\varphi_{\max} \ .\end{aligned}$$

where (a) follows since $\int_{\mathcal{S}} \chi_i(s)ds = 1$. Now consider the case where $\varphi_{\max} > 0$. Then as γ goes to 1, we have that

$$\int_{\mathcal{N}_\gamma} \chi_i(s)ds = \int_{\mathcal{N}_\gamma} \psi_i(s)\mu(s)ds \to 0$$

since $\mu[\mathcal{N}_\gamma] \to 0$ as $\gamma \to 1$. Let $\gamma_0 \in [0,1]$ be sufficiently close to 1 such that for all i,

$$c_i \coloneqq \int_{\mathcal{N}_{\gamma_0}} \chi_i(s)ds \leq \frac{1}{2} \ .$$

Then we have

$$\begin{aligned}\int_S \chi_i(s)\varphi(s)ds &= \int_{\mathcal{N}_{\gamma_0}} \chi_i(s)\varphi(s)ds + \int_{S\setminus\mathcal{N}_{\gamma_0}} \chi_i(s)\varphi(s)ds \\ &\overset{(a)}{\leq} \int_{\mathcal{N}_{\gamma_0}} \chi_i(s)\varphi_{\max}ds + \int_{S\setminus\mathcal{N}_{\gamma_0}} \chi_i(s)\gamma_0\varphi_{\max}ds \\ &\overset{(b)}{=} \varphi_{\max}\big(c_i + \gamma_0(1-c_i)\big) \\ &\overset{(c)}{\leq} \frac{1}{2}\big(1+\gamma_0\big)\varphi_{\max}\end{aligned}$$

where (a) follows since for any $s \notin \mathcal{N}_{\gamma_0}$, $\varphi(s) < \gamma_0\varphi_{\max}$. (b) follows by definition of c_i above and (c) follows since c_i is upper bounded by $\frac{1}{2}$. So for $\gamma = \frac{1}{2}\big(1+\gamma_0\big)$, $\int_{s\in S} \chi_i(s)\varphi(s)ds \leq \gamma\varphi_{\max}$.

Now suppose φ has a non-flat minimum. If $\varphi_{\min} \geq 0$, then $\gamma\varphi_{\min} < \varphi_{\min}$ for all $\gamma \in (0,1)$, and thus

$$\begin{aligned}\int_{\mathcal{S}} \chi_i(s)\varphi(s)ds &\geq \varphi_{\min} \int_{\mathcal{S}} \chi_i(s)ds \\ &= \varphi_{\min} \\ &\geq \gamma\varphi_{\min} \ .\end{aligned}$$

Now consider the case where $\varphi_{\min} < 0$. Let γ_0 be such that

$$c_i \coloneqq \int_{\mathcal{N}_{\gamma_0}^-} \chi_i(s)ds \leq \frac{1}{2}$$

Then

$$\begin{aligned}\int_S \chi_i(s)\varphi(s)ds &= \int_{\mathcal{N}_{\gamma_0}^-} \chi_i(s)\varphi(s)ds + \int_{S\setminus\mathcal{N}_{\gamma_0}^-} \chi_i(s)\varphi(s)ds \\ &\overset{(a)}{\geq} c_i\varphi_{\min} + \gamma_0\varphi_{\min}(1-c_i) \\ &\overset{(b)}{\geq} \frac{1}{2}\big(1+\gamma_0\big)\varphi_{\min}\end{aligned}$$

where (a) follows by definition of c_i, and (b) follows since c_i is upper bounded by $\frac{1}{2}$ and $\varphi_{\min}$ is a negative value. Thus for $\gamma = \frac{1}{2}\big(1+\gamma_0\big)$, $\int_S \chi_i(s)\varphi(s)ds \geq \gamma\varphi_{\min}$.

Combining the bounds in both cases, we have for $\gamma \geq \frac{1}{2}\big(1+\gamma_0\big)$ that

$$\gamma\varphi_{\min} \leq \int_{\mathcal{S}} \chi_i(s)\varphi(s)ds \leq \gamma\varphi_{\max}\ .$$

From theorem 5.18, this implies that ambiguity holds. □

5.3.3 Ambiguity for general Ψ

In this section we derive a sufficient condition under which *any* natural algorithm faces ambiguity. In paricular, we do not assume that assumption 7 holds. Our only assumption is the following

Assumption 8. Let $\Psi = \{\psi_1, \ldots, \psi_k\}$ be a basis for $im(\Pi_\Psi)$. Also assume that $\|\psi_i\|_{1,\mu} = \int_{\mathcal{S}} \psi_i(s)\mu(s)ds = 1$ for all i.

Remark 5.22. We note that such a basis always exists for natural projection operators and as discussed in assumption 7, requiring $\|\psi_i\|_{1,\mu} = 1$ is not restrictive. Thus, our results under assumption 8 show ambiguity holds for any natural projection operator and hence, for any natural algorithm.

Before we show our next result, we first show that P_T is a non-expansion.

Proposition 5.23. *The operator P_T is a non-expansion w.r.t. $\|\cdot\|_\mu$.*

Proof. Let $S' = S$. For $V \in L_2(\mathcal{S}, \mu)$, we have

$$\begin{aligned}
\|P_T V\|_\mu^2 &= \langle P_T V, P_T V \rangle_\mu \\
&= \int_{\mathcal{S}} \mu(s)(P_T V(s))^2 ds \\
&= \int_{\mathcal{S}} \mu(s) \left(\int_{\mathcal{S}'} T(s'|s) V(s') ds' \right)^2 ds \\
&\overset{(a)}{\leq} \int_{\mathcal{S}} \mu(s) \int_{\mathcal{S}'} T(s'|s) V(s')^2 ds' ds \\
&\overset{(b)}{=} \int_{\mathcal{S}'} \int_{\mathcal{S}} \mu(s) T(s'|s) V(s')^2 ds ds' \\
&\overset{(c)}{=} \int_{\mathcal{S}'} \mu(s') V(s')^2 ds' \\
&= \left\| V(s') \right\|_\mu^2
\end{aligned}$$

where (a) follows by Jensen's inequality, (b) follows by the Tonelli-Fubini theorem, and (c) follows since μ is the stationary distribution. Since the quadratic function is monotonically increasing on $\mathbb{R}_+$, we have that $\|P_T V\|_\mu \leq \left\| V(s') \right\|_\mu$ and so P_T is non-expansive. □

The next result helps provide a useful re-definition of corollary 5.15 that will aid in the main result we show next.

Theorem 5.24. *Assume that assumptions 6 and 8. Ambiguity holds if and only if there exists a function $f : S \to [\varphi_{\min}, \varphi_{\max}] \in L^2(S, \mu)$ such that for all i,*

$$\int_{s \in S} \chi_i(s) \varphi(s) ds = \gamma \int_{s \in S} \chi_i(s) f(s) ds \ .$$

Proof. By corollary 5.15, ambiguity holds if and only if there exists a T such that for all i

$$\int_{s \in S} \chi_i(s) \varphi(s) ds = \gamma \int_{s \in S} \chi_i(s) P_T \varphi(s) ds$$

By proposition 5.23, P_T is a non-expansion. Thus, $\varphi_{\min} \leq P_T \varphi(s) \leq \varphi_{\max}$. Then let $f(s) \coloneqq P_T \varphi(s)$ and our result holds. □

We now use theorem 5.24 to prove our main result in the case of a continuous state space.

Theorem 5.25. *Assume that assumptions 6 and 8. Suppose there exists a $\varphi \neq 0 \in im(\Pi_\Psi)$ that has non-flat maximum and minimum. Then ambiguity holds.*

Proof. We look to show that there exists a function $f : \mathcal{S} \to [\varphi_{\min}, \varphi_{\max}]$ that satisfies

$$\int_{s\in S} \chi_i(s)\varphi(s)ds = \gamma \int_{s\in S} \chi_i(s)f(s)ds \ , \tag{5.4}$$

and thus, ambiguity holds by theorem 5.24. We first look to find a function f that satisfies equation (5.4) and whose range is upper bounded by $\varphi_{\max}$. Consider the function $f(s) = \frac{1}{\gamma}\varphi(s)$. Clearly f satisfies equation (5.4). Also, for $\varphi_{\max} < 0$, f satisfies the upper bound since $\frac{1}{\gamma}\varphi(s) \leq \varphi(s) \leq \varphi_{\max}$. However, for $\varphi_{\max} > 0$, f exceeds the upper bound. We now look to construct a function f that satisfies the upper bound for all values of $\varphi_{\max}$. Let

$$\bar{f}(s) \coloneqq \frac{1}{\gamma}\varphi(s) - \delta(s) \ , \tag{5.5}$$

where $\delta(s) \coloneqq \max\left\{0, \frac{1}{\gamma}\varphi(s) - \gamma\varphi_{\max}\right\}$. By construction, $\bar{f}$ satisfies the upper bound. Now let us define

$$f(s) \coloneqq \bar{f}(s) + g(s) \ .$$

We now construct g such that f satisfies the equation 5.4. The following derivation derives a linear system of equations that constrain the function g such that f satisfies equation (5.4). Starting from equation (5.4), we have

$$\begin{aligned}\int_S \chi_i(s)\varphi(s)ds &= \gamma \int_S \chi_i(s)f(s)ds \\ &\overset{(a)}{=} \gamma \int_S \chi_i(s)\left(\frac{1}{\gamma}\varphi(s) - \delta(s) + g(s)\right)ds \ .\end{aligned}$$

Here (a) follows by the definition of f and $\bar{f}$. Cancelling out $\int_S \chi_i(s)\varphi(s)ds$ and γ from both sides of the equation gives

$$\int_S \chi_i(s)\delta(s)ds = \int_S \chi_i(s)g(s)ds \ . \tag{5.6}$$

We now note that the set of functions $\chi_1, \ldots, \chi_k$ forms a linearly independent set. To see this, suppose that for all $s \in \mathcal{S}$

$$b_1\chi_1(s) + b_2\chi_2(s) + \ldots + b_k\chi_k(s) = 0$$

and $b_1, \ldots b_k$ are not all equal to 0. Then since $\chi_i(s) = \mu(s)\psi_i(s)$, and $\mu(s) > 0$, we must have

$$b_1\psi_1(s) + b_2\psi_2(s) \ldots + b_k\psi_k(s) = 0 \ .$$

This is a contradiction since the set of functions $\Psi = \{\psi_1, \ldots, \psi_k\}$ is a linearly independent set. Thus, let g be given by

$$g(s) = \sum_{j=1}^{k} \chi_j(s) a_j \ , \forall s \in \mathcal{S}$$

where $a_i \in \mathbb{R}$ for all i. Then from equation 5.6 we have

$$\begin{aligned} \int_S \chi_i(s)\delta(s)ds &= \int_{\mathcal{S}} \chi_i(s) \sum_{j=1}^{k} \chi_j(s) a_j ds \\ &\overset{(a)}{=} \sum_{j=1}^{k} a_j \int_{\mathcal{S}} \chi_i(s)\chi_j(s)ds \end{aligned}$$

where in (a) we swapped the summation and the integrand by Tonelli-Fubini theorem. Now as a notational shorthand, let $\langle f, g \rangle \coloneqq \int_{\mathcal{S}} f(s)g(s)ds$. Then we have

$$\sum_{j=1}^{k} a_j \int_{\mathcal{S}} \chi_i(s)\chi_j(s)ds = \sum_{j=1}^{k} a_j \langle \chi_i, \chi_j \rangle \ .$$

Let $X \in \mathbb{R}^{k \times k}$ and $\bar{\delta} \in \mathbb{R}^k$ be defined by

$$\begin{aligned} X_{ij} &= \langle \chi_i, \chi_j \rangle \ , \quad i, j = 1, \ldots, k \\ \bar{\delta}_i &= \langle \chi_i, \delta \rangle \ , \quad i = 1, \ldots, k \end{aligned}$$

respectively. Together a, X and $\bar{\delta}$ form a system of linear equations given by

$$Xa = \bar{\delta}$$

Note that since $\chi_1, \ldots, \chi_k$ is a set of linearly independent functions, X is full rank and thus invertible. We can express a as

$$a = X^{-1}\bar{\delta} \ .$$

Let $\chi(s) = \big(\chi_1(s), \ldots, \chi_d(s)\big)$. We can now express g as

$$g(s) = \chi(s)X^{-1}\bar{\delta} \ , \quad \forall s \in \mathcal{S} \ .$$

We now explicitly define three infinity norms we will use to bound g. For a function $f(s) = (f_1(s), \ldots, f_k(s))$, vector $u \in \mathbb{R}^k$, and matrix $A \in \mathbb{R}^{k \times k}$, the

norms are given by

$$\begin{aligned}
\|f\|_\infty &\coloneqq \max_{1\le i\le k} \sup_{s\in\mathcal{S}} \left|f_i(s)\right| \ , \\
\|u\|_\infty &\coloneqq \max_{1\le i\le k} |u_i| \ , \\
\|A\|_\infty &\coloneqq \sup_{y\neq 0} \frac{\|Ay\|_\infty}{\|y\|_\infty} = \max_j \sum_{i=1}^{k} \left|A_{ij}\right| \ .
\end{aligned}$$

Under these norm definitions, we see that $\left\|X^{-1}\right\|_\infty < \infty$ since X is invertible. The function χ has its infinity norm given by $\|\chi\|_\infty = \max_{1\le i\le k} \sup_{s\in\mathcal{S}} \left|\chi_i(s)\right|$. To see that this is finitely bounded, recall that for any i, $\int_{\mathcal{S}} \chi_i(s)ds = 1$. We can then split the χ_i into two functions χ_i^+ and χ_i^- where χ_i^+ is the same value as χ_i when it is positive and χ_i^- is the same value as χ_i when it is negative. Then since

$$\int_{\mathcal{S}} \chi_i(s)ds = \int_{\mathcal{S}} \chi_i^+(s) - \int_{\mathcal{S}} \chi_i^-(s)ds = 1 \ ,$$

it must be the case that both individual integrals are finite. Thus $\left|\chi_i(s)\right| < \infty$ for all i and s. We now look to derive a bound for $\bar{\delta}$. Let

$$\mathcal{N}_{\gamma^2} \coloneqq \left\{ s \in \mathcal{S} : \varphi(s) \ge \gamma^2 \varphi_{\max} \right\} \ .$$

Note that $\delta(s) > 0$ if and only if $s \in \mathcal{N}_{\gamma^2}$. Let $\mathbf{1}_{\mathcal{N}_{\gamma^2}}$ be the characteristic function for $\mathcal{N}_{\gamma^2}$. Then for all $i = 1, \ldots, k$, we have the following derivation

$$\begin{aligned}
\int_{\mathcal{S}} \chi_i(s)\delta(s)ds &\overset{(a)}{=} \int_{\mathcal{S}} \psi_i(s)\delta(s)\mu(s)ds \\
&\overset{(b)}{\le} \left(\frac{1}{\gamma}\varphi_{\max} - \gamma\varphi_{\max}\right) \int_{\mathcal{S}} \mathbf{1}_{\mathcal{N}_{\gamma^2}}(s)\psi_i(s)\mu(s)ds \\
&\overset{(c)}{=} \left(\frac{1}{\gamma}\varphi_{\max} - \gamma\varphi_{\max}\right) \psi_i(s_{\max}) \int_{\mathcal{N}_{\gamma^2}} \mu(s)ds \\
&\overset{(d)}{=} \left(\frac{1}{\gamma}\varphi_{\max} - \gamma\varphi_{\max}\right) \psi_i(s_{\max})\mu\left[\mathcal{N}_{\gamma^2}\right] \ .
\end{aligned}$$

Here (a) follows by definition of χ_i and (b) follows by definition of δ. In (c), we let $s_{\max}$ denote the value of $s \in \mathcal{S}$ that ψ_i achieves a maximum. Finally, (d) follows by definition of $\mu\left[\mathcal{N}_{\gamma^2}\right]$. Thus, $\bar{\delta}$ can be bounded by

$$\left\|\bar{\delta}\right\|_\infty = \max_{1\le i\le k} \left|\bar{\delta}\right| \le \varphi_{\max} \left(\frac{1}{\gamma} - \gamma\right) \psi_i(s_{\max})\mu\left[\mathcal{N}_{\gamma^2}\right] \ .$$

Combining the bounded quantities, we have that g is bounded by

$$\begin{aligned}\|g\|_\infty &\leq \|\chi\|_\infty \left\|X^{-1}\right\|_\infty \varphi_{\max}\left(\frac{1}{\gamma}-\gamma\right)\psi_i(s_{\max})\mu\left[\mathcal{N}_{\gamma^2}\right] \\ &\leq C\cdot\varphi_{\max}\left(\frac{1}{\gamma}-\gamma\right)\mu\left[\mathcal{N}_{\gamma^2}\right]\end{aligned}$$

where to simplify notation we let C be the constant defined as $C = \|\chi\|_\infty \left\|X^{-1}\right\|_\infty \psi_i(s_{\max})$. Note that by definition, $\bar{f}$ is bounded by $\gamma\varphi_{\max}$. As a result, we can now bound f from above by

$$\begin{aligned}f(s) &= \bar{f}(s) + g(s) \\ &\leq \gamma\varphi_{\max} + \|g\|_\infty \\ &\leq \varphi_{\max}\left(\gamma + C\left(\frac{1}{\gamma}-\gamma\right)\mu\left[\mathcal{N}_{\gamma^2}\right]\right) \ .\end{aligned}$$

Now $\varphi_{\max}\left(\gamma + C\left(\frac{1}{\gamma}-\gamma\right)\mu\left[\mathcal{N}_{\gamma^2}\right]\right) \leq \varphi_{\max}$ if and only if

$$\gamma + C\left(\frac{1}{\gamma}-\gamma\right)\mu\left[\mathcal{N}_{\gamma^2}\right] \leq 1 \ . \tag{5.7}$$

This preceding inequality holds if and only if

$$\mu\left[\mathcal{N}_{\gamma^2}\right] \leq \frac{1-\gamma}{C\left(\frac{1}{\gamma}-\gamma\right)} \ .$$

Now

$$\begin{aligned}\frac{1-\gamma}{C\left(\frac{1}{\gamma}-\gamma\right)} &= \frac{\gamma-\gamma^2}{C\left(1-\gamma^2\right)} \\ &= \frac{\left(\gamma-\gamma^2\right)(1+\gamma)}{C(1-\gamma)^2(1+\gamma)} \\ &= \frac{\gamma-\gamma^3}{C(1-\gamma)^2(1+\gamma)} \\ &= \frac{\gamma}{C\left(1+\gamma\right)} \ .\end{aligned}$$

Hence equation (5.7) holds if and only if $\mu\left[\mathcal{N}_{\gamma^2}\right] \leq \frac{\gamma}{C(1+\gamma)}$. Since φ has a non-flat maximum, by lemma 5.20 we have that as $\gamma \to 1$

$$\mu\left[\mathcal{N}_{\gamma^2}\right] \to 0 \ ,$$

whilst we have as $\gamma \to 1$,

$$\frac{\gamma}{C\,(1+\gamma)} \to \frac{1}{2C}\ .$$

Thus, $\mu\left[\mathcal{N}_{\gamma^2}\right] \leq \frac{\gamma}{C(1+\gamma)}$ if φ has non-flat maximum and γ is sufficiently close to 1. Thus $f(s) \leq \varphi_{\max}$ for all $s \in \mathcal{S}$ and f satisfies the upper bound. For more intuition, please see Figure 5.3 for an impression of the idea underlying this construction.

By a similar argument to the above, we can also find a function f^- that satisfies equation (5.4) and has range greater than the lower bound given φ has non-flat minimum. If $\varphi_{\min} \geq 0$, then $f^-(s) = \frac{1}{\gamma}\varphi(s) \geq \varphi_{\min}$ and equation (5.4). We thus consider the case where $\varphi_{\min} < 0$. We define f^- by

$$f^-(s) = \tilde{f}(s) - g^-(s)\ , \forall s \in \mathcal{S}$$

where g^- is some function chosen such that f^- satisfies equation (5.4). The function $\tilde{f}(s)$ is given by

$$\tilde{f}(s) = \frac{1}{\gamma}\varphi(s) + \delta^-(s)\ ,$$

where $\delta^-(s)$ is given by

$$\delta^-(s) = \max\{0, \gamma\varphi_{\min} - \frac{1}{\gamma}\varphi(s)\}\ .$$

By construction, $\tilde{f}$ satisfies the lower bound. We now look to derive a system of linear equations to constrain g^- such that equation (5.4) is satisfied. Starting from equation (5.4) we have

$$\begin{aligned}\int_{\mathcal{S}} \chi_i(s)\varphi(s)ds &= \gamma \int_{\mathcal{S}} \chi_i(s) f^-(s) ds \\ &= \gamma \int_{\mathcal{S}} \chi_i(s) \left(\frac{1}{\gamma}\varphi(s) + \delta^-(s) - g^-(s)\right) ds\ .\end{aligned}$$

Cancelling out from both sides $\int_{\mathcal{S}} \chi_i(s)\varphi(s)$ ds and re-arranging gives

$$\int_{\mathcal{S}} \chi_i(s)\delta^-(s)ds = \int_{\mathcal{S}} \chi_i(s) g^-(s) ds\ .$$

We let g^- be given by

$$g^-(s) = \sum_{j=1}^{k} \chi_j(s) a_j\ .$$

Let $\bar{\delta}^- \in \mathbb{R}^k$ be given by

$$\bar{\delta}^- = \langle \chi_i, \delta^- \rangle \ , \quad i = 1, \ldots, k \ .$$

Then g^- is given by

$$g^-(s) = \chi(s) X^{-1} \delta^- \ .$$

We now look to bound $\bar{\delta}^-$. For all $i = 1, \ldots, k$ we have the following derivation

$$\begin{aligned}
\int_{\mathcal{S}} \chi_i(s)\delta^-(s)ds &= \int_{\mathcal{S}} \psi_i(s)\delta^-(s)\mu(s)ds \\
&\overset{(a)}{\leq} \varphi_{\min}\left(\gamma - \frac{1}{\gamma}\right) \int_{\mathcal{N}_{\gamma^2}^-} \psi_i(s)\mu(s)ds \\
&\overset{(b)}{\leq} \varphi_{\min}\left(\gamma - \frac{1}{\gamma}\right) \psi_i(s_{\min}) \int_{\mathcal{N}_{\gamma^2}^-} \mu(s)ds \\
&\overset{(c)}{=} \varphi_{\min}\left(\gamma - \frac{1}{\gamma}\right) \psi_i(s_{\min}) \mu\left[\mathcal{N}_{\gamma^2}^-\right] \ .
\end{aligned}$$

In (a), we used the fact that $\delta^-(s)$ is only greater than 0 for $s \in \mathcal{N}_{\gamma^2}^-$ and that it is upper bounded by $\varphi_{\min}\left(\gamma - \frac{1}{\gamma}\right)$. In (b) we define $s_{\min}$ as the value of $s \in \mathcal{S}$ where ψ_i achieves a minimum. Finally, (c) follows by definition of $\mu\left[\mathcal{N}_{\gamma^2}^-\right]$. Having bound $\int_{\mathcal{S}} \chi_i(s)\delta^-(s)ds$ for all i, we have that $\bar{\delta}^-$ is bound in the infinity norm

$$\bar{\delta}^- \leq \varphi_{\min}\left(\gamma - \frac{1}{\gamma}\right) \psi_i(s_{\min}) \mu\left[\mathcal{N}_{\gamma^2}^-\right] \ .$$

Thus, $g^-(s)$ is bounded in the infinity norm by

$$\begin{aligned}
\left\|g^-\right\|_\infty &\leq \|\chi\|_\infty \left\|X^{-1}\right\|_\infty \left\|\delta^-\right\|_\infty \\
&= D\varphi_{\min}\left(\gamma - \frac{1}{\gamma}\right) \mu\left[\mathcal{N}_{\gamma^2}^-\right]
\end{aligned}$$

where to simplify notation we define the constant $D \coloneqq \|\chi\|_\infty \left\|X^{-1}\right\|_\infty \psi_i(s_{\min})$. Given that for any $s \in \mathcal{S}$ $\tilde{f}(s)$is bound below by $\gamma\varphi_{\min}$, we have that $f^-(s)$ is bounded as follows.

$$\begin{aligned}
f^-(s) &= \tilde{f}^-(s) - g^-(s) \\
&\geq \gamma\varphi_{\min} - D\varphi_{\min}\left(\gamma - \frac{1}{\gamma}\right) \mu\left[\mathcal{N}_{\gamma^2}^-\right] \ .
\end{aligned}$$

Now $\gamma\varphi_{\min} - D\varphi_{\min}\left(\gamma - \frac{1}{\gamma}\right)\mu\left[\mathcal{N}_{\gamma^2}^-\right] \geq \varphi_{\min}$ if and only if

$$\gamma + D\left(\frac{1}{\gamma} + \gamma\right)\mu\left[\mathcal{N}_{\gamma^2}^-\right] \leq 1 \ .$$

since $\varphi_{\min} < 0$. Re-arranging shows that this inequality holds if and only if

$$\begin{aligned}\mu\left[\mathcal{N}_{\gamma^2}^-\right] &\leq \frac{1-\gamma}{D\left(\frac{1}{\gamma} - \gamma\right)} \\ &= \frac{\gamma}{D\left(1+\gamma\right)} \ .\end{aligned}$$

Since φ has non-flat minimum, we have that $\mu\left[\mathcal{N}_{\gamma^2}^-\right] \to 0$ as $\gamma \to 1$ whereas $\frac{\gamma}{D(1+\gamma)} \to \frac{1}{2D}$. Thus, the inequality holds if φ has non-flat minimum and γ is sufficiently close to 1.

We now combine the two cases into one function. Let f be given by

$$f(s) \coloneqq \frac{1}{\gamma}\varphi(s) - \delta(s) + g(s) + \delta^-(s) - g^-(s) \ .$$

We note that the two sets $\mathcal{N}_{\gamma^2}$ and $\mathcal{N}_{\gamma^2}^-$ do not intersect and so the two cases do not overlap. Thus under this definition, $\varphi_{\min} \leq f(s) \leq \varphi_{\max}$ for all $s \in S$ and f satisfies the constraint presented in equation (5.4) if φ has non-flat maximum and minimum. Thus for any Ψ ambiguity holds if there exists a $\varphi \neq 0$ with non-flat maximum and minimum. □

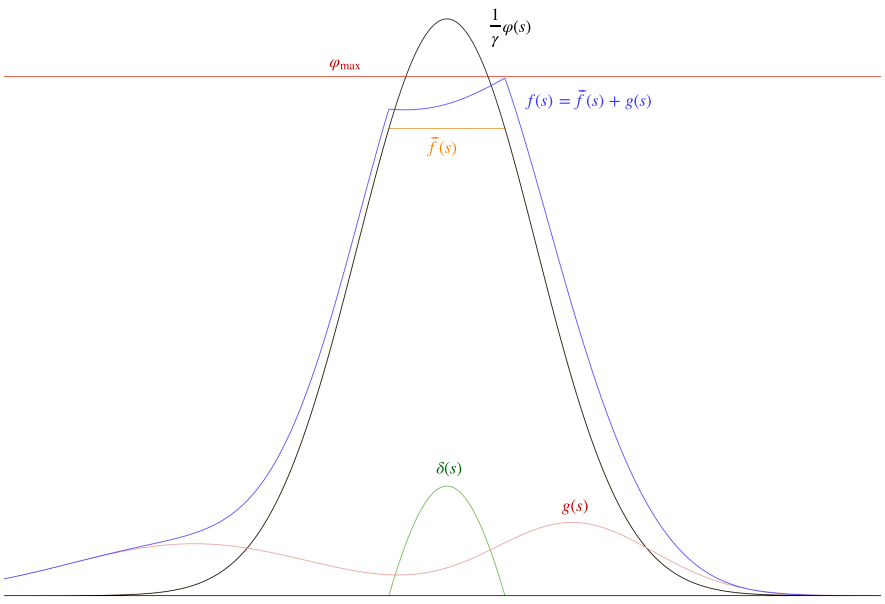

Figure 5.3: An impression of our construction of f in Theorem 5.25 that satisfies the upper bound.

5.4 Summary

In summary, we have seen in this chapter different conditions under which natural algorithms may experience a subtle form of divergence which we call *ambiguity*; under ambiguity, natural algorithms may either diverge or converge to the wrong solution. Under these conditions, natural algorithms exhibit a more subtle form of divergence: either they diverge or they converge to the wrong solution. When the state space is finite, we showed generic conditions under which ambiguity holds. For a particular projection, namely when the image of its transpose has an orthonormal basis where each element is also a positive vector, we were able to characterise ambiguity in terms of the error in the value function. We were also able to use these conditions to explicitly construct a simple example where there are multiple environments that 'look' the same under projection, that is have the same quantities A, B, and b, whilst each having very different parameter vectors. When the state space is a compact subspace of $\mathbb{R}$, we were able to show that the conditions that held in the finite case held analogously. Furthermore, we were able to show a sufficient condition under which ambiguity holds when any projection is taken on the Bellman equations.

Chapter 6

Extensions and Future Outlook

Having concluded our main results in the last chapter, we now discuss some possible extensions to our results and the open problems that remain.

Of mathematical interest is extending our results to more general linear fixed point equations. In [2], Bertsekas describes a general linear fixed point equation as

$$x = b + Ax$$

where $b \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$. An interesting question is whether a projected linear fixed point equation also has key quantities that characterise its solution as well as whether or not our ambiguity results can generalise for methods that look to solve these more general linear fixed point equations.

In relation to reinforcement learning the most pertinent task is to explicitly determine whether other linear function approximation methods fall in our class of natural algorithms or whether our results extend to more methods that take projections on 'Bellman-like' equations. The most immediate set of algorithms to consider are the $TD(\lambda)$ algorithms for $\lambda > 0$. Recall from theorem [reference] $TD(\lambda)$'s solution is characterised as the fixed point of $\Pi T^{(\lambda)}$, where Π is the orthogonal projection and $T^{(\lambda)}$ is the $TD(\lambda)$ operator, which is 'Bellman-like'. We suspect that our results should easily extend for values of λ close to 0, since these methods would behave more similarly to $TD(0)$, but it remains an open question. We also conjecture that the gradient temporal difference algorithm (GTD for short) presented in [18] is a natural algorithm. One possible line of analysis is to consider whether both GTD and $TD(\lambda)$ compute statistical estimates that are related to A, B and b in some fashion, similar to how $LSTD$ behaves.

Another immediate task is to analyse algorithms that do not fit as natural algorithms and are guaranteed to converge and compare for key differences. In

particular, the emphatic temporal difference algorithm, presented in [19], is one such algorithm to investigate.

Finally, one other natural extension is to consider how our results can extend to algorithms that operate using the Q-value function instead of the value function. We suspect that there may be easy extensions to defining a class of natural Q-learning algorithms, of which least-squares Q-learning [23] and Q-learning [22] may be candidate members.

Thus, it is clear that a plethora of open problems exist to provide extensions to the results shown in this thesis.

Appendix A

Appendices

A.1 Monte-Carlo Methods

As described by Halton [10], a Monte-Carlo method "represent[s] the solution of a problem as a parameter of a hypothetical population, and use[s] a random sequence of numbers to construct a sample of the population, from which statistical estimates of the population can be obtained". Its theoretical validation depends heavily upon the law of large numbers. However, the point here will not be to rigorously justify Monte-Carlo simulation but to appeal to the reader's intuition. We now provide a simple example to illustrate this phenomenon.

Computing the value of π using Monte-Carlo simulation

This example is drawn from [13]. Suppose we wish to compute the value of π using random samples. Assume that we can produce sample points within the square $[-1,1] \times [-1,1]$ such that the probability of a sample point falling within some region $\mathcal{R} \subset [-1,1]^2$ is proportional to the area of $\mathcal{R}$ but independent of the position of $\mathcal{R}$. In this case, it is clear to see that the two coordinates X, Y are distributed as uniform random variables over the interval $[-1,1]$. Now suppose $\mathcal{R}$ is the unit circle. Then the probability that a random sample point falls within $\mathcal{R}$ is given by

$$\mathbb{P}(\text{point within } \mathcal{R}) = \frac{\text{area of circle}}{\text{area of } [-1,1]^2} = \frac{\iint_{x^2+y^2 \leq 1} 1 dxdy}{\iint_{-1 \leq x,y \leq 1} 1 dxdy} = \frac{\pi}{4} \ .$$

Thus, we can express the value of π in terms of the probability of a point falling within $\mathcal{R}$.

$$\pi = 4 \cdot \mathbb{P}(\text{point within } \mathcal{R}) \ .$$

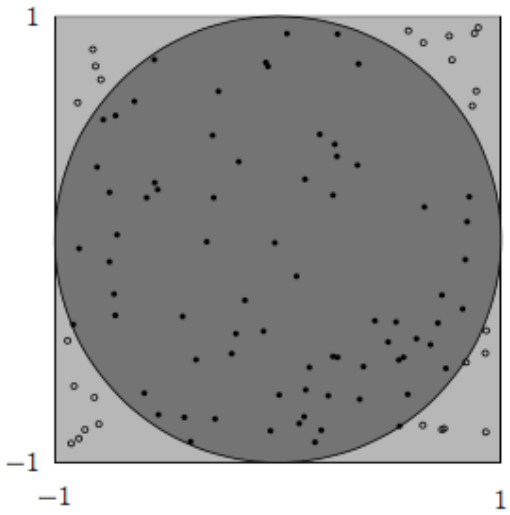

Figure A.1: Illustration of the random sampling of points to estimate the value of π. The image is drawn from [13].

The following graph shows the estimation of π as the number of random samples generated increases.

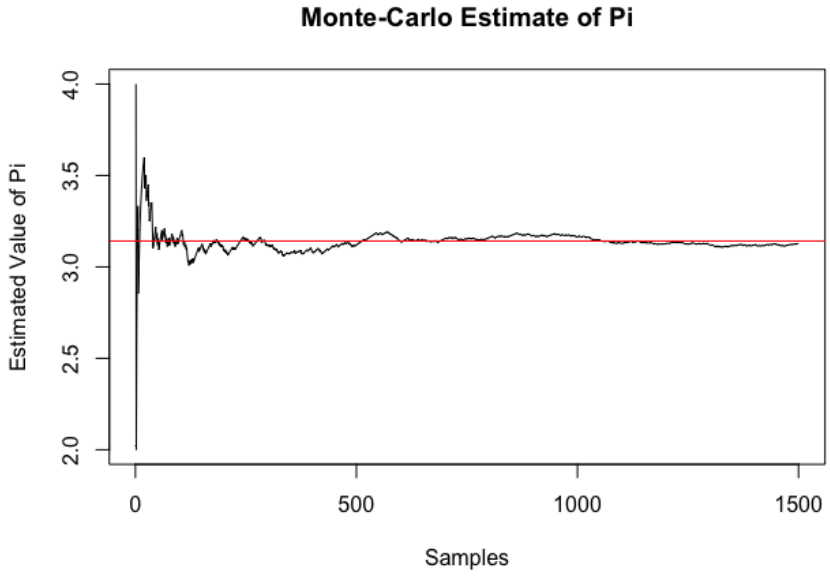

Figure A.2: Graph of the results of Monte-Carlo simulation to estimate π over 1500 samples. The red line indicates the true value of π.

Bibliography

[1] L. Baird. Residual algorithms: Reinforcement learning with function approximation. *International Conference on Machine Learning*, 1995.

[2] D. P. Bertsekas. *Dynamic Programming and Optimal Control 3rd Edition, Volume II.* Massachusetts Institute of Technology, 2011.

[3] D. P. Bertsekas and J. Tsitsiklis. *Neuro-Dynamic Programming.* Massachusetts Institute of Technology, 1996.

[4] A. Borovkov. *Probability Theory.* Springer-Verlag London, 2013.

[5] S. Bradke and A. Barto. Linear least-squares algorithms for temporal difference learning. *Machine Learning*, 22:33–57, 1996.

[6] J. Conway. *A Course in Functional Analysis.* Springer-Verlag New York, Inc., 1990.

[7] A. Eberle. Markov processes. `https://wt.iam.uni-bonn.de/fileadmin/WT/Inhalt/people/Andreas_Eberle/Markov_Processes_1617/MPSkript1617.pdf`.

[8] H. Georgii. *Stochastics.* Walter de Gruyter GmbH, 2013.

[9] Gockenbach. *Understanding and Implementing the Finite Element Method.* Society for Industrial and Applied Mathematics, 2006.

[10] J. Halton. A retrospective and prospective survey of the monte carlo method. *SIAM Review*, 12:1–63, 1970.

[11] M. Hutter. (im)possibility of linear reinforcement learning. Technical Report.

[12] Q. Jin. Numerical optimisation. Mathematical Sciences Institute, The Australian Natural University, Lecture Notes 2018.

[13] A. Johansen and L. Evers. Monte-carlo methods. University of Bristol, Dept. of Mathematics, Lecture Notes 2007.

[14] P. Portal. Mathematics of finance. Mathematical Sciences Institute, The Australian National University, Lecture Notes 2017.

[15] B. Scherrer. Should one compute the temporal difference fix point or minimize the bellman residual ? the unified oblique projection view. arXiv:1011.4362.

[16] R. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3:9–44, 1988.

[17] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 2018.

[18] R. Sutton, H. Maei, and C. Szepesvár. A convergent o(n) temporal-difference algorithm for off-policy learning with linear function approximation. *Advances in Neural Information Processing Systems 21*, 2009.

[19] R. Sutton, A. Mahmood, and M. White. An emphatic approach to the problem of off-policy temporal-difference learning. arXiv:1503.04269.

[20] C. Szepesvári. *Algorithms for Reinforcement Learning*. Morgan and Claypool Publishers, 2010.

[21] J. Tsitsiklis and B. V. Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997.

[22] C. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8(3):279292, 1992.

[23] H. Yu and D. Bertsekas. A least squares q-learning algorithm for optimal stopping problems. *LIDS REPORT 2731*, 2007.