ABSTRACT

Title of dissertation: LEARNING ALGORITHMS FOR MARKOV DECISION PROCESSES

Abraham Thomas, Doctor of Philosophy, 2009

Dissertation directed by: Professor Steven Marcus
Department of Electrical and Computer Engineering

We propose various computational schemes for solving Partially Observable Markov Decision Processes with the finite stage additive cost and infinite horizon discounted cost criterion. Error bounds for the corresponding algorithms are given and it is further shown that at the expense of more computational effort the Partially Observable Markov Decision Problem (POMDP) can be solved as closely to the optimal as desired.

It is well known that a sufficient statistic for taking the best action at any time for the POMDP is the aposteriori probability distribution on the underlying states, given all the past history, and that this can be updated recursively. We prove that the finite stage optimal costs as well as the optimal cost for the infinite horizon discounted cost problem are both Lipschitz continuous (with domain the unit simplex of probability distributions over the underlying states) and gives bounds for the Lipschitz constant. We use these bounds to provide error bounds for computational algorithms for solving POMDPs.

We extend the almost sure convergence result of a very general stochastic approximation algorithm to the case when the underlying Markov process exhibits periodicity. This result is used to extend the proof of convergence of Temporal Difference (TD) reinforcement learning schemes with linear function approximation for Markov Cost processes in order to estimate the cost to go function for the discounted cost criterion, and the differential cost function for the average cost criterion, respectively.

Adaptive control of Markov Decision Problems (MDPs) is a problem in which a full knowledge of the system parameters, namely transition probabilities as well as the distribution of the immediate costs, are not available apriori. We give direct adaptive control schemes for infinite horizon discounted cost and average cost MDPs. Approximate Policy Iteration using on-line TD schemes for policy evaluation is detailed for the discounted cost and average cost criteria.

Possible extensions of direct adaptive control schemes to the POMDP framework are discussed.

Auxiliary results relevant to the core results of the dissertation are stated and proved in the appendices. In particular an efficient discretization scheme for the finite dimensional unit simplex is given. Some general error bounds for MDPs are also given. Also TD schemes for learning in Stochastic Shortest Path problems (SSP) are discussed.

LEARNING ALGORITHMS FOR MARKOV DECISION PROCESSES

by

Abraham Thomas

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2009

Advisory Committee:
Professor Steven Marcus, Chair/Advisor
Professor Mark Shayman
Professor Prakash Narayan
Professor Armand Makowski
Professor Mike Boyle

Acknowledgments

I owe my gratitude to all the people who have made this dissertation possible and because of whom my graduate experience has been one that I will cherish forever.

First and foremost I'd like to thank my advisor, Professor Steven Marcus for giving me an invaluable opportunity to work on challenging and extremely interesting problems during my doctoral studies. I am extremely grateful for his able guidance and patience during the course of my doctoral studies. It has been a pleasure to work with and learn from such an extraordinary individual.

I would also like to thank Professor Mark Shayman, Professor Prakash Narayan, Professor Armand Makowski and Professor Mike Boyle for agreeing to serve on my dissertation committee and for sparing their invaluable time reviewing the manuscript.

I would like to acknowledge financial support from the Institute for Systems Research and the Electrical and Computer Engineering Department at the University of Maryland, College Park.

I would also like to acknowledge the help and support from the staff members at the ECE Graduate Office, ISR, IES and Graduate School Office of the University of Maryland, College Park.

Table of Contents

Chapter 1

Introduction

In this dissertation we propose direct adaptive control schemes for Markov Decision Processes (MDPs) and suggest their extension for Partially Observable Markov Decision Processes (POMDPs). We also consider some discretization schemes for solving POMDPs approximately.

In this chapter, we define the basic finite state, finite action Markov Decision Process model as well as the finite state, finite action, finite observation Partially Observable Markov Decision Process model. We also introduce a standard Stochastic Approximation Algorithm, which can be used to prove the convergence of Temporal Difference schemes for evaluating the cost to go function for the infinite horizon discounted cost criterion and the differential cost function for the average cost criterion respectively of Markov Cost processes.

Subsequently we give short introductions to the contents of each of the following chapters as well as the appendices. This essentially sums up the contributions of the dissertation.

1.1 Markov Decision Processes

A Markov Decision Process (MDP) [12, 40] is a system which evolves as follows. Let $\mathbf{N}_0$ denote the set of whole numbers and $\mathbf{R}$ denote the set of real numbers. At

any discrete time $t \in \mathbf{N}_0$, the state of the system is $s_t \in \mathcal{S}$, where $\mathcal{S}$ is the set of possible states or state space. While in state s_t we can execute one out of a set $\mathcal{A}(s_t)$ of feasible actions. The state space as well as the feasible action set for each state are assumed to be non-empty. Define $\mathcal{A} = \bigcup_{i \in \mathcal{S}} \mathcal{A}(i)$ as the action space. Upon execution of an action $u_t \in \mathcal{A}(s_t)$ at time t, the system moves to state s_{t+1} at time $t+1$ and an immediate cost $g_t \in \mathbf{R}$ (which may be random but which depends on s_t, u_t and s_{t+1}) is incurred. The new state s_{t+1} occurs with a probability which depends on s_t and u_t. Given s_t and u_t the state transition probability distribution of s_{t+1} does not depend on the past values of states, actions or immediate costs. Similarly given s_t, u_t and s_{t+1}, the probability distribution of the immediate cost g_t also does not depend on the past values of states, actions or immediate costs. This is essentially the Markov property of the problem. Let $h_t = (s_0, u_0, g_0, s_1, u_1, g_1, \ldots, s_{t-1}, u_{t-1}, g_{t-1}, s_t)$ denote the history of the process upto time t with $h_0 = (s_0)$. The history follows the recursion $h_t = (h_{t-1}, u_{t-1}, g_{t-1}, s_t)$ for $t \geq 1$.

An admissible policy ν is a sequence of stochastic kernels $\{\nu_t\}$ on $\mathcal{A}$ given the past history h_t, with the restriction that $\nu_t(\mathcal{A}(s_t) \mid h_t) = 1$, that is, the probability measure is concentrated on the set of feasible actions. Note that $\nu = \{\nu_0, \nu_1, \nu_2, \ldots\}$.

In this dissertation we focus our attention primarily on finite state, finite action homogeneous MDPs where the state space and action space (along with the feasible action sets) does not change over time, nor do the the state transition probabilities and the distribution of the immediate cost. For convenience we denote $\mathcal{S} \equiv \{1, 2, \ldots, n\}$ and $\mathcal{A}(i) \equiv \{1, 2, \ldots, \mid \mathcal{A}(i) \mid\}$, for $i \in \mathcal{S}$. The state space and feasible action sets for each state are non-empty. Here $\mid \mathcal{A}(i) \mid$ denotes the cardinality

of the set $\mathcal{A}(i)$. Now $|\mathcal{S}|$ and $|\mathcal{A}|$ are finite numbers. Here $\mathcal{A} = \bigcup_{i\in\mathcal{S}} \mathcal{A}(i)$. The transition probabilities may be conveniently denoted by $p_{ij}(u) = \Pr[s_{t+1} = j \mid s_t = i, u_t = u]$, where $i, j \in \mathcal{S}$ and $u \in \mathcal{A}(i)$. Here Pr denotes probability. For $u \in \mathcal{A}(i)$, let $g(i, u, j)$ denote the expected value $\mathrm{E}[g_t \mid s_t = i, u_t = u, s_{t+1} = j]$. Then the expected value of the immediate cost for taking action $u \in \mathcal{A}(i)$ from state i is $g(i, u) \equiv \mathrm{E}[g_t \mid s_t = i, u_t = u] = \sum_{j=1}^{n} p_{ij}(u)g(i, u, j)$. We assume these expectations to be finite. $\mathcal{S}$ and $\mathcal{A}$ are endowed with the discrete topology. $\mathbf{R}$ is endowed with the Borel topology. Let $\mathcal{H}_t$ denote the set of all histories up to time t. Here $\mathcal{H}_0 = \mathcal{S}$, $\mathcal{H}_t = \mathcal{H}_{t-1}\mathcal{A}\mathbf{R}\mathcal{S}$. These spaces are endowed with the corresponding product topologies. Here $\Omega = \mathcal{H}^\infty = (\mathcal{S}\mathcal{A}\mathbf{R})^\infty$ is the sample space under consideration. $\mathcal{H}^\infty$ is the set of infinite sequences of the form $(s_0, u_0, g_0, s_1, u_1, g_1, \ldots)$ where $s_t \in \mathcal{S},\ u_t \in \mathcal{A}$ and $g_t \in \mathbf{R}$.

The set of all admissible policies is denoted by $\mathcal{M}$ (the set of history dependent randomized policies). A policy ν is said to be Markov if ν_t depends only on the current state s_t and t and not on the past history, that is $\nu_t(\cdot \mid h_t) = \delta_t(\cdot \mid s_t)$, where δ_t is a stochastic control kernel which takes a probability distribution on $\mathcal{A}(i)$ for each state $i \in \mathcal{S}$. To be precise, it is called a Markov randomized policy. If all the probability mass is concentrated on a single action for each $i \in \mathcal{S}$, we call it a Markov deterministic policy. In this case we may think of control functions μ_t on $\mathcal{S}$ with $\mu_t(i) \in \mathcal{A}(i)$, instead of the stochastic kernel δ_t. A Markov randomized policy is said to be stationary if $\delta_t = \delta$ for all $t \in \mathbf{N}_0$. For convenience we denote such a stationary randomized policy with δ. If we have a Markov deterministic policy in which $\mu_t = \mu$ for all $t \in \mathbf{N}_0$, we call it a stationary deterministic policy. For convenience we

denote such a policy with μ. The set of all (Markov) stochastic control kernels or, equivalently, all the stationary randomized policies is denoted by Λ. A (Markov) stochastic control kernel δ may be denoted as follows. $\delta(i)$ represents a probability distribution on the set $\mathcal{A}(i)$ for each $i \in \mathcal{S}$. $[\delta(i)]_a$ represents the probability of executing action $a \in \mathcal{A}(i)$ from state $i \in \mathcal{S}$. $[\delta(i)]_a \geq 0$ and $\sum_{a \in \mathcal{A}(i)} [\delta(i)]_a = 1$. Likewise the set of all control functions or equivalently stationary deterministic policies is denoted by Υ. $\mu \in \Upsilon$ iff $\mu(i) \in \mathcal{A}(i), \ \forall i \in \mathcal{S}$. The cardinality of Υ is given by $|\Upsilon| = \prod_{i=1}^{n} |\mathcal{A}(i)|$. For a measure theoretic approach to MDPs with general state and action spaces please refer to [4, 14, 23].

1.1.1 Cost Criterion

MDPs may be classified on the basis of the cost structure we try to minimize. Let $\mathcal{P}_i^{\nu}(\cdot) \equiv \mathcal{P}^{\nu}(\cdot \mid s_0 = i)$ denote the probability distribution induced on Ω under the policy ν, when we start from state $s_0 = i$. $\mathrm{E}^{\nu}[\cdot \mid s_0 = i]$ denotes the corresponding expectation. We are concerned only with variations of additive cost problems.

In the *finite horizon* problem we try to minimize

$$\mathrm{E}^{\nu}\left[\sum_{t=0}^{N-1} \beta^t g_t + \beta^N G(s_N) \mid s_0 = i\right]$$

for each $i \in \mathcal{S}$. Let $\mathbf{N}$ denote the set of natural numbers. Here $N \in \mathbf{N}$ is the horizon and $\beta^N G(s_N)$ is the terminal cost incurred for being in state s_N at time N where $\beta \in [0, \infty)$. The expectation is with respect to the probability measure induced by the policy ν. Note that for the N stage problem, only $\nu_0, \ \nu_1, \ \ldots, \ \nu_{N-1}$ are relevant in the computation of the expectation.

In the *infinite horizon discounted cost* criterion we try to minimize

$$\mathrm{E}^{\nu}\left[\sum_{t=0}^{\infty}\beta^{t}g_{t}\mid s_{0}=i\right]$$

for each $i\in\mathcal{S}$. Here $\beta\in[0,1)$ is the discount factor. This quantity is well defined and is equal to

$$\lim_{N\to\infty}\mathrm{E}^{\nu}\left[\sum_{t=0}^{N-1}\beta^{t}g_{t}\mid s_{0}=i\right]$$

Here the costs incurred in the future are given less weight because of the discount factor.

In the *average cost* formulation we try to minimize

$$\limsup_{N\to\infty}\frac{1}{N}\mathrm{E}^{\nu}\left[\sum_{t=0}^{N-1}g_{t}\mid s_{0}=i\right]$$

for each $i\in\mathcal{S}$.

Yet another cost formulation is the *stochastic shortest path* formulation where we try to minimize the total cost

$$\limsup_{N\to\infty}\mathrm{E}^{\nu}\left[\sum_{t=0}^{N-1}g_{t}\mid s_{0}=i\right]$$

for each $i\in\{1,2,\ldots,n\}$. Here we assume that there is an additional state 0, which is a cost free termination state; once the system reaches that state it remains there at no further cost (i.e. zero cost). The structure of the problem is assumed to be such that termination is inevitable, at least under an optimal policy. Thus the objective is to reach the termination state with minimal expected cost. The problem is in effect a finite horizon problem, but the length of the horizon may be random and may be affected by the policy being used. We may assume WLOG that there

is only one feasible action at state 0, namely action 1 (i.e. $\mathcal{A}(0) = \{1\}$), under which the system remains at state 0, incurring an immediate cost of zero. That is $\mathrm{E}[|g_t| \mid s_t = 0, u_t = 1] = 0$ and $p_{00}(1) = \Pr[s_{t+1} = 0 \mid s_t = 0, u_t = 1] = 1$. With $g(i,u,j) \equiv \mathrm{E}[g_t \mid s_t = i, u_t = u, s_{t+1} = j]$, $g(i,u) \equiv \mathrm{E}[g_t \mid s_t = i, u_t = u] = \sum_{j=0}^{n} g(i,u,j)$ for $i, j \in \{0, 1, \ldots, n\}$, $u \in \mathcal{A}(i)$,

We have the following important lemma from [40, Theorem 5.5.1].

Lemma 1.1 *Let $\nu = \{\nu_0, \nu_1, \nu_2, \ldots\}$ be any history dependent randomized policy. Then for each fixed $i \in \mathcal{S}$, there exists a Markov randomized policy $\nu' = \{\delta_0, \delta_1, \delta_2, \ldots\}$ dependent on i and ν such that*

$$\mathcal{P}^{\nu}(s_t = j, u_t = a \mid s_0 = i) = \mathcal{P}^{\nu'}(s_t = j, u_t = a \mid s_0 = i)$$

for all $t \in \mathbf{N}_0$, $j \in \mathcal{S}$, $a \in \mathcal{A}(j)$. Also

$$\mathcal{P}^{\nu}(s_t = j \mid s_0 = i) = \mathcal{P}^{\nu'}(s_t = j \mid s_0 = i)$$

□

Notice that we may choose

$$[\delta_t(j)]_a = \mathcal{P}^{\nu}(u_t = a \mid s_t = j, s_0 = i)$$

for $t \in \mathbf{N}_0$, $j \in \mathcal{S}$, $a \in \mathcal{A}(j)$. Here $\mathcal{P}^{\nu}$ and $\mathcal{P}^{\nu'}$ denote the probability measures induced by policy ν and ν' respectively.

1.1.2 Optimality Criterion

1.1.2.1 Finite Horizon Problem

We now state the Dynamic Programming (DP) Algorithm [11] for the homogeneous finite horizon problem. For every initial state i, the optimal cost $J^*(i)$ of the basic problem is equal to $J_0(i)$, where the function $J_0 \in \mathbf{R}^n$ is given by the last step of the following algorithm (value iteration), which proceeds backward in time from stage $N-1$ to stage 0:

$$\begin{aligned} J_N(i) &= G(i), \qquad i \in \mathcal{S} \\ J_k(i) &= \min_{u \in \mathcal{A}(i)} \left[g(i,u) + \beta \sum_{j=1}^{n} p_{ij}(u) J_{k+1}(j) \right], \qquad i \in \mathcal{S} \\ & \quad k = 0, \ldots, N-1 \end{aligned}$$

Let μ_k^* be the control function such that $\mu_k^*(i)$ is a minimizing action in the above equation. The N stage policy $\nu^* = \{\mu_0^*, \ldots, \mu_{N-1}^*\}$ is optimal for the N-stage problem. Note that the above computation easily extends to the non-homogeneous MDP, though we are concerned mostly with homogeneous MDPs.

1.1.2.2 Infinite Horizon Discounted Cost Problem

For the infinite horizon discounted cost problem with discount factor $\beta \in [0,1)$, we denote by $J^\nu \in \mathbf{R}^n$ the cost to go vector associated with following policy $\nu \in \mathcal{M}$ and is given by

$$J^\nu(i) = \mathrm{E}^\nu \left[\sum_{t=0}^{\infty} \beta^t g_t \mid s_0 = i \right], \qquad i \in \mathcal{S}.$$

For the infinite horizon discounted cost problem define for each deterministic control function $\mu \in \Upsilon$, the following operator $T_\mu : \mathbf{R}^n \to \mathbf{R}^n$, by

$$(T_\mu J)(i) = g(i, \mu(i)) + \beta \sum_{j=1}^{n} p_{ij}(\mu(i)) J(j)$$

for each $J \in \mathbf{R}^n$. In vector notation $T_\mu J = \bar{g}^\mu + \beta P_\mu J$, where $\bar{g}^\mu \in \mathbf{R}^n$ is the expected immediate cost vector for policy μ, with $\bar{g}^\mu(i) = g(i, \mu(i))$ and P_μ is the $n \times n$ transition probability matrix with $[P_\mu]_{ij} = p_{ij}(\mu(i))$.

Similarly, define the dynamic programming operator [12] $T : \mathbf{R}^n \to \mathbf{R}^n$ as follows

$$(TJ)(i) = \min_{u \in \mathcal{A}(i)} \left[g(i, u) + \beta \sum_{j=1}^{n} p_{ij}(u) J(j) \right].$$

We may use the following vector notation, namely $TJ = \min_{\mu \in \Upsilon} T_\mu J$, where the minimization is componentwise. Note that $\tilde{\mu} = \arg\min_{\mu \in \Upsilon} T_\mu J$ iff $T_{\tilde{\mu}} J = TJ$. It is possibile that there may be more than one minimizing control function. It may be easily seen that both the operator T as well as T_μ are monotone, i.e. if $J, \tilde{J} \in \mathbf{R}^n$ with $J \leq \tilde{J}$, then $TJ \leq T\tilde{J}$ and $T_\mu J \leq T_\mu \tilde{J}$. Here the inequality is componentwise, i.e. $J \leq \tilde{J}$ means $J(i) \leq \tilde{J}(i)$ for each $i \in \{1, \ldots, n\}$. Also they have the property that

$$T(J + \alpha \underline{1}) = T(J) + \beta\, \alpha \underline{1}$$

and

$$T_\mu(J + \alpha \underline{1}) = T_\mu(J) + \beta\, \alpha \underline{1}$$

for $\alpha \in \mathbf{R}$ and any stationary deterministic policy $\mu \in \Upsilon$. Here $\underline{1}$ is the vector in $\mathbf{R}^n$ with all components equal to one. Hence it is easy to see that T and T_μ are

contraction mappings under the supremum norm $\| \cdot \|$ with contraction coefficient β. That is

$$\| TJ - T\tilde{J} \| \leq \beta \| J - \tilde{J} \|$$

and

$$\| T_\mu J - T_\mu \tilde{J} \| \leq \beta \| J - \tilde{J} \|$$

for $J, \tilde{J} \in \mathbf{R}^n$. Here for $J \in \mathbf{R}^n$, the supremum norm (or sup-norm) is given by

$$\| J \| = \max_{1 \leq i \leq n} |J(i)|$$

The contraction mappings T and T_μ have unique fixed points. That is, there exists $J^* \in \mathbf{R}^n$ such that

$$TJ^* = J^* \tag{1.1}$$

and $J^\mu \in \mathbf{R}^n$ such that

$$T_\mu J^\mu = J^\mu.$$

In fact, it can be shown [12, 23] that J^* is the optimal cost to go function (or vector) for the infinite horizon discounted cost problem, and J^μ is the cost to go function (or vector) associated with following the stationary policy μ. That is,

$$J^*(i) = \inf_{\nu \in \mathcal{M}} \mathrm{E}^\nu \left[\sum_{t=0}^{\infty} \beta^t g_t \mid s_0 = i \right]$$

and

$$J^\mu(i) = \mathrm{E}^\mu \left[\sum_{t=0}^{\infty} \beta^t g_t \mid s_0 = i \right],$$

for $i \in \mathcal{S}$. Equation 1.1 is called the Bellman equation for the discounted cost problem. It can be shown that a stationary deterministic policy $\tilde{\mu} \in \Upsilon$ is optimal

iff

$$\tilde{\mu}(i) = \arg \min_{u \in \mathcal{A}(i)} \left[g(i,u) + \beta \sum_{j=1}^{n} p_{ij}(u) J^*(j) \right]$$

for all $i \in \mathcal{S}$. In fact, it can also be shown that

$$\begin{aligned} J^\mu &= \sum_{k=0}^{\infty} \beta^t P_\mu^k \, \bar{g}^\mu \\ &= (I - \beta P_\mu)^{-1} \bar{g}^\mu, \end{aligned}$$

where $P_\mu^0 \equiv I$, is the identity matrix. P_μ^k is P_μ raised to the k^{th} power.

Define, for each $\delta \in \Lambda$, the expected immediate cost vector $\bar{g}^\delta \in \mathbf{R}^n$ as $\bar{g}^\delta(i) = \sum_{a \in \mathcal{A}(i)} [\delta(i)]_a g(i,a)$ and the $n \times n$ transition probability matrix P_δ to be $[P_\delta]_{ij} = \sum_{a \in \mathcal{A}(i)} [\delta(i)]_a p_{ij}(a)$. Consider the operator $T_\delta : \mathbf{R}^n \to \mathbf{R}^n$ given by

$$T_\delta J = \bar{g}^\delta + \beta \, P_\delta J$$

for $J \in \mathbf{R}^n$. In fact

$$\left(T_\delta J\right)(i) = \sum_{a \in \mathcal{A}(i)} [\delta(i)]_a \left[g(i,a) + \beta \sum_{j=1}^{n} p_{ij}(a) J(j) \right]$$

for $i \in \mathcal{S}$. T_δ is also a monotone operator which is a contraction mapping under the sup-norm with contraction coefficient β. It has a unique fixed point J^δ. The cost to go vector corresponding to the stationary policy δ is given by $J^\delta = \sum_{t=0}^{\infty} \beta^t P_\delta^t \bar{g}^\delta = (I - \beta P_\delta)^{-1} \bar{g}^\delta$. Any $\delta \in \Lambda$ is optimal iff $T_\delta J^* = T J^*$.

1.1.2.3 Average Cost Problem

Note that for any policy $\nu \in \mathcal{M}$, the average cost vector $\bar{\vartheta}^\nu \in \mathbf{R}^n$ denotes the average cost to go function, namely

$$\bar{\vartheta}^\nu(i) = \limsup_{N \to \infty} \frac{1}{N} \mathrm{E}^\nu \left[\sum_{t=0}^{N-1} g_t \mid s_0 = i \right]$$

for each $i \in \mathcal{S}$. For stationary policies the limit exists [12], i.e. for any stationary policy $\delta \in \Lambda$

$$\bar{\vartheta}^{\delta}(i) = \lim_{N\to\infty} \frac{1}{N} \mathrm{E}^{\delta} \left[\sum_{t=0}^{N-1} g_t \mid s_0 = i \right]$$

for $i \in \mathcal{S}$.

Let $\bar{\vartheta}^* \in \mathbf{R}^n$ denote the optimal average cost vector given by

$$\bar{\vartheta}^*(i) = \inf_{\nu\in\mathcal{M}} \bar{\vartheta}^{\nu}(i)$$

for $i \in \mathcal{S}$. We add that [12, 40]

$$\begin{aligned} \bar{\vartheta}^*(i) &= \inf_{\nu\in\mathcal{M}} \liminf_{N\to\infty} \frac{1}{N} \mathrm{E}^{\nu} \left[\sum_{t=0}^{N-1} g_t \mid s_0 = i \right] \\ &= \inf_{\nu\in\mathcal{M}} \limsup_{N\to\infty} \frac{1}{N} \mathrm{E}^{\nu} \left[\sum_{t=0}^{N-1} g_t \mid s_0 = i \right]. \end{aligned} \tag{1.2}$$

Let $\bar{g}^{\delta}$, the expected immediate cost vector and P_{δ}, the transition probability matrix corresponding to stationary randomized policy δ, be defined as earlier. Note that

$$\bar{\vartheta}^{\delta} = \left(\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} P_{\delta}^{k} \right) \bar{g}^{\delta}.$$

An important result regarding transition probability matrices is that the limit in the preceding equation exists [12].

For $\delta \in \Lambda$, define the operator $\bar{T}_{\delta} : \mathbf{R}^n \to \mathbf{R}^n$ by

$$\bar{T}_{\delta} J = \bar{g}^{\delta} + P_{\delta} J$$

for $J \in \mathbf{R}^n$. Define the operator $\bar{T} : \mathbf{R}^n \to \mathbf{R}^n$ by

$$\bar{T} J = \min_{\mu\in\Upsilon} \bar{T}_{\mu} J$$

for $J \in \mathbf{R}^n$. Here, minimization is done componentwise, namely

$$(\bar{T}J)(i) = \min_{a \in \mathcal{A}(i)} \left[g(i,a) + \sum_{j=1}^{n} p_{ij}(a) J(j) \right].$$

For the average cost problem we have the following result [12]. If a scalar $\vartheta \in \mathbf{R}$ and a vector $J \in \mathbf{R}^n$ satisfy

$$\vartheta \underline{1} + J = \bar{T}J,$$

then ϑ is the optimal average cost per stage $\bar{\vartheta}^*(i)$ for all $i \in \mathcal{S}$. Furthermore, if $\bar{T}_{\delta^*} J = \bar{T}J$, then the stationary policy δ^* is optimal, i.e. $\bar{\vartheta}^{\delta^*} = \vartheta \underline{1} = \bar{\vartheta}^*$. Also, if the optimal average cost starting from any state is the same, namely $\vartheta^* \in \mathbf{R}$, then there exists $J \in \mathbf{R}^n$ such that [12, Proposition 4.1.4]

$$\vartheta^* \underline{1} + J = \bar{T}J.$$

We have the following corollary. Let δ be a stationary policy. If a scalar ϑ and a vector $J \in \mathbf{R}^n$ satisfy

$$\vartheta \underline{1} + J = \bar{T}_\delta J$$

then $\bar{\vartheta}^\delta(i) = \vartheta, \ \forall i \in \mathcal{S}$. Infact if $\bar{\vartheta}^\delta(i) = \vartheta^\delta, \ \forall i \in \mathcal{S}$, then there exists $J \in \mathbf{R}^n$ such that

$$\vartheta^\delta \underline{1} + J = \bar{T}_\delta J.$$

A stationary deterministic policy μ^* is said to be *Blackwell optimal* if it is simultaneously optimal for all β-discounted infinite horizon problems with β in an interval $(\bar{\beta}, 1)$, where $\bar{\beta}$ is some scalar with $0 < \bar{\beta} < 1$. For the finite state, finite action MDP there exists a Blackwell optimal policy. Blackwell optimal policies are

average cost optimal, irrespective of whether the optimal average cost is the same for all starting states $i \in \mathcal{S}$ [12, 40].

1.1.2.4 Stochastic Shortest Path Problem

The cost to go function (or vector) for the stochastic shortest path problem (SSP) corresponding to policy $\nu \in \mathcal{M}$ is given by

$$\tilde{J}^{\nu}(i) = \limsup_{N \to \infty} \mathrm{E}^{\nu} \left[\sum_{t=0}^{N-1} g_t \mid s_0 = i \right] \tag{1.3}$$

for $i \in \{1, 2, \ldots, n\}$. For the SSP we say that a stationary deterministic policy $\mu \in \Upsilon$ (assume WLOG that in the termination state 0 we take the unique feasible action namely 1, under which the system remains in state 0 at zero cost) is proper if when using this policy, there is a positive probability that the termination state will be reached after at most n stages, regardless of the initial state, i.e.

$$\max_{i \in \{1,2,\ldots,n\}} \mathcal{P}^{\mu} \left(s_n \neq 0 \mid s_0 = i \right) < 1.$$

A similar definition of properness exists for stationary randomized policies. A stationary policy that is not proper is called improper. For a stationary proper policy, the limit exists in equation 1.3, i.e. lim sup may be replaced by lim. For the SSP, define for each stationary deterministic policy $\mu \in \Upsilon$, the $n \times n$ sub-stochastic matrix P_{μ} to be $[P_{\mu}]_{ij} = p_{ij}(\mu(i))$ for $i, j \in \{1, 2, \ldots, n\}$. Similarly, define the expected immediate cost vector $\bar{g}^{\mu} \in \mathbf{R}^n$ to be $\bar{g}^{\mu}(i) = g(i, \mu(i))$ for $i \in \{1, 2, \ldots, n\}$. Note that $g(i, u) \equiv \mathrm{E}\left[g_t \mid s_t = i, u_t = u \right] = \sum_{j=0}^{n} p_{ij}(u) g(i, u, j)$ for $i \in \{1, 2, \ldots, n\},\ u \in \mathcal{A}(i)$. Here,

$$g(i, u, j) \equiv \mathrm{E}\left[g_t \mid s_t = i, u_t = u, s_{t+1} = j \right]$$

for $i, j \in \{0, 1, 2, \ldots, n\},\ u \in \mathcal{A}(i)$. Similarly, for a stationary randomized policy $\delta \in \Lambda$, define the $n \times n$ sub-stochastic matrix P_δ by

$$[P_\delta]_{ij} = \sum_{a \in \mathcal{A}(i)} [\delta(i)]_a p_{ij}(a)$$

for $i, j \in \{1, 2, \ldots, n\}$. Define the expected immediate cost vector $\bar{g}^\delta \in \mathbf{R}^n$ by $\bar{g}^\delta(i) = \sum_{a \in \mathcal{A}(i)} [\delta(i)]_a g(i, a)$ for $i \in \{1, \ldots, n\}$.

Define for stationary policy $\delta \in \Lambda$, the operator $\tilde{T}_\delta : \mathbf{R}^n \to \mathbf{R}^n$ by

$$\tilde{T}_\delta J = \bar{g}^\delta + P_\delta J$$

for $J \in \mathbf{R}^n$. Also define the operator $\tilde{T} : \mathbf{R}^n \to \mathbf{R}^n$ by $\tilde{T}J = \min_{\mu \in \Upsilon} \tilde{T}_\mu J$, where the minimization is componentwise. That is,

$$\left(\tilde{T}J\right)(i) = \min_{a \in \mathcal{A}(i)} \left[g(i, a) + \sum_{j=1}^{n} p_{ij}(a) J(j) \right]$$

for $J \in \mathbf{R}^n$ and $i \in \{1, 2, \ldots, n\}$. The cost to go function (or vector) for the SSP corresponding to a stationary policy $\delta \in \Lambda$ is given by

$$\tilde{J}^\delta(i) = \limsup_{N \to \infty} \left[\sum_{k=0}^{N-1} P_\delta^k \bar{g}^\delta \right]_i$$

for $i \in \{1, 2, \ldots, n\}$.

We make the following assumptions [12].

Assumption 1.1 *There exists at least one stationary deterministic proper policy.*

Assumption 1.2 *For every improper stationary deterministic policy μ, the corresponding cost $\tilde{J}^\mu(i)$ is ∞ for at least one state $i \in \{1, 2, \ldots, n\}$, i.e. some component of the sum $\sum_{k=0}^{N-1} P_\mu^k \bar{g}^\mu$ diverges to ∞ as $N \to \infty$.*

A stationary deterministic policy $\mu \in \Upsilon$ satisfying, for some vector $J \in \mathbf{R}^n$, the relation $\tilde{T}_\mu J \leq J$ (the inequality is componentwise) is proper under Assumption 1.1 and Assumption 1.2 [12].

Under Assumption 1.1 and Assumption 1.2 the optimal cost to go vector $\tilde{J}^* \in \mathbf{R}^n$ is the unique solution of Bellman's equation

$$\tilde{T}\tilde{J}^* = \tilde{J}^*.$$

Here,

$$\tilde{J}^*(i) = \inf_{\mu \in \mathcal{M}} \tilde{J}^\nu(i)$$

for $i \in \{1, 2, \ldots, n\}$. A stationary deterministic policy $\mu \in \Upsilon$ is optimal iff

$$\tilde{T}_\mu \tilde{J}^* = \tilde{T}\tilde{J}^*.$$

Note that such a μ is proper. For a proper policy $\delta \in \Lambda$, the cost to go vector is given by

$$\tilde{J}^\delta = \lim_{N \to \infty} \sum_{k=0}^{N-1} P_\delta^k \bar{g}^\delta = (I - P_\delta)^{-1} \bar{g}^\delta.$$

1.2 Value Iteration and Policy Iteration

In this section we will discuss the two main schemes for solving the MDPs (that is finding the optimal cost to go and optimal policies). We will be discussing *Value Iteration* and *Policy Iteration* for infinite horizon discounted cost problems and SSPs. The value iteration schemes and policy iteration schemes for the general average cost problem are more involved and will not be discussed here. See [12, 40] for details. In this dissertation we are interested in average cost policy iteration

schemes for unichain [12, 40] MDPs with a common recurrent state. This will be discussed in Chapter 5.

1.2.1 Value Iteration

First we focus on the infinite horizon discounted cost problem. Since T is a contraction mapping with contraction coefficient β, we have the result that for any two vectors $J, \hat{J} \in \mathbf{R}^n$ and for all $k = 0, 1, \ldots$, there holds

$$\max_{i \in \mathcal{S}} |(T^k J)(i) - (T^k \hat{J})(i)| \leq \beta^k \max_{i \in \mathcal{S}} |J(i) - \hat{J}(i)|.$$

In the value iteration scheme we start with any vector $J \in \mathbf{R}^n$ and successively compute $TJ,\ T^2J, \ldots$. Here $T^kJ = T(T^{k-1}J)$ for $k \in \mathbf{N}$ with $T^0J = J$. Since T is a contraction mapping, we have [12]

$$\lim_{k \to \infty} (T^k J)(i) = J^*(i)$$

for all $i \in \mathcal{S}$. Here, J^* is the optimal cost to go function for the infinite horizon discounted cost problem. Furthermore, the error sequence $|(T^kJ)(i) - J^*(i)|$ is bounded by a constant multiple of β^k for all $i \in \mathcal{S}$. This method is also called successive approximation.

For the SSP we assume that Assumption 1.1 and Assumption 1.2 hold. The DP operator $\tilde{T}$ is in general not a contraction mapping. In the value iteration scheme we start with a vector $J \in \mathbf{R}^n$ and successively compute $\tilde{T}J,\ \tilde{T}^2J, \ldots$. Here $\tilde{T}^kJ = \tilde{T}(\tilde{T}^{k-1}J)$ for $k \in \mathbf{N}$, with $\tilde{T}^0J = J$. Under Assumption 1.1 and Assumption 1.2 we have [12]

$$\lim_{k \to \infty} (\tilde{T}^k J)(i) = \tilde{J}^*(i)$$

for all $i \in \{1, 2, \ldots, n\}$. Here, $\tilde{J}^* \in \mathbf{R}^n$ is the optimal cost to go function for the SSP. This method is also called successive approximation.

1.2.2 Policy Iteration

The policy iteration algorithm generates a sequence of stationary deterministic policies, each with improved cost over the preceding one.

First we deal with the discounted cost problem. Given the stationary deterministic policy μ, and the corresponding cost function J^μ, an improved policy $\bar{\mu}$ is computed by minimization in the Dynamic Programming (DP) equation corresponding to J^μ, that is $T_{\bar{\mu}} J^\mu = T J^\mu$, and the process is repeated. The algorithm is based on the following result [12].

Let μ and $\bar{\mu}$ be stationary deterministic policies such that $T_{\bar{\mu}} J^\mu = T J_\mu$, or equivalently, for $i = 1, \ldots, n$,

$$g(i, \bar{\mu}(i)) + \beta \sum_{j=1}^{n} p_{ij}(\bar{\mu}(i)) J^\mu(j) = \min_{u \in \mathcal{A}(i)} \left[g(i, u) + \beta \sum_{j=1}^{n} p_{ij}(u) J^\mu(j) \right].$$

Then we have

$$J^{\bar{\mu}}(i) \leq J^\mu(i), \qquad i = 1, \ldots, n.$$

Furthermore, if μ is not optimal, strict inequality holds in the above equation for at least one state i.

The policy iteration algorithm is given below.

Step 1: (Initialization) Guess an initial stationary deterministic policy μ_0.

Step 2: (Policy Evaluation) Given the stationary deterministic policy μ_k, compute

the corresponding cost function J^{μ_k} from the linear system of equations

$$(I - \beta P_{\mu_k}) J^{\mu_k} = \bar{g}^{\mu_k}.$$

Step 3: (Policy Improvement) If $J^{\mu_k} = TJ^{\mu_k}$ stop; else obtain a new stationary deterministic policy μ_{k+1} satisfying

$$T_{\mu_{k+1}} J^{\mu_k} = TJ^{\mu_k}$$

and go to step 2 and repeat the process.

□

Note that since the number of stationary deterministic policies is finite, policy iteration algorithm converges in a finite ($\leq |\Upsilon|$) steps.

Now we discuss the policy iteration scheme for SSP. The policy iteration for SSP is along the same lines as for the discounted cost problem. The policy iteration algorithm generates a sequence of proper stationary deterministic policies, each with improved cost over the preceding one. Given a proper stationary deterministic policy μ and the corresponding cost to go function $\tilde{J}^{\mu} \in \mathbf{R}^n$, an improved proper stationary deterministic policy $\bar{\mu}$ is obtained by minimization in the DP equation corresponding to $\tilde{J}^{\mu}$, that is $\tilde{T}_{\bar{\mu}} \tilde{J}^{\mu} = \tilde{T} \tilde{J}^{\mu}$, and the process is repeated. The algorithm is based on the following result [12]. Let μ be proper stationary deterministic policy. Let $\bar{\mu}$ be a stationary deterministic policy such that $\tilde{T}_{\bar{\mu}} \tilde{J}^{\mu} = \tilde{T} \tilde{J}^{\mu}$ or equivalently

$$g(i, \bar{\mu}(i)) + \sum_{j=1}^{n} p_{ij}(\bar{\mu}(i)) \tilde{J}^{\mu}(j) = \min_{u \in \mathcal{A}(i)} \left[g(i,u) + \sum_{j=1}^{n} p_{ij}(u) \tilde{J}^{\mu}(j) \right].$$

Then $\bar{\mu}$ is a proper policy and

$$\tilde{J}^{\bar{\mu}}(i) \leq \tilde{J}^{\mu}(i), \qquad i = 1, \ldots, n.$$

Furthermore if μ is not optimal, strict inequality holds in the above equation for at least one state i.

The policy iteration algorithm for SSP is as in the discounted problem. We start with a proper stationary deterministic policy μ_0. In step 2 (policy evaluation) we compute the cost to go function by

$$\tilde{J}^{\mu_k} = (I - P_{\mu_k})^{-1} \bar{g}^{\mu_k}.$$

For asynchronous value iteration, modified policy iteration, and approximate policy iteration see [12, 16]. For adaptive aggregation schemes see [12, 13]. For parallel distributed implementations see [15].

1.3 Partially Observable Markov Decision Processes

These are problems in which we cannot directly observe the current state of the process for decision making [5, 6, 35, 39, 47, 48, 49] . Instead we get noisy observations of the underlying state transitions. In this problem we assume that the feasible control actions for all the underlying states are the same, namely $\mathcal{A}$. Here at time t, the system is in state $s_t \in \mathcal{S}$, but we don't have access to this state information. We take an action $u_t \in \mathcal{A}$ and the system moves to state s_{t+1} with probability $p_{s_t s_{t+1}}(u_t)$, incurs a cost g_t with $\mathrm{E}[g_t \mid s_t = i, u_t = u, s_{t+1} = j] = g(i, u, j)$ and $\mathrm{E}[g_t \mid s_t = i, u_t = u] = g(i, u)$. An observation $y_{t+1} \in \mathcal{O} = \{1, 2, \ldots, |\mathcal{O}|\}$ is observed with probability $Q(y_t \mid s_t, u_t, s_{t+1})$. This additional information can be utilized for taking an action at time $t + 1$. We deal with finite state, finite action, finite observation POMDPs. A sufficient statistic for taking the best action at any

time t is the aposteriori probability distribution on the underlying states given the history of past actions and observations and the initial distribution on the underlying states. This aposteriori probability may be computed recursively at each time step. A more systematic approach to the definition of the POMDP is given in Chapter 2. Here again we may have different cost criteria like finite horizon, infinite horizon discounted cost and average cost formulation. We will be primarily interested in finite horizon and infinite horizon discounted cost criteria.

1.4 A Stochastic Approximation Algorithm

Next we consider a stochastic approximation algorithm [8] which is used in proving the convergence of temporal difference schemes [54, 55]

Consider the following algorithm

$$\theta_{t+1} = \theta_t + \gamma_{t+1} H(\theta_t, X_{t+1}) + \gamma_{t+1}^2 \rho_{t+1}(\theta_t, X_{t+1}),$$

where θ_t evolves in $\mathbf{R}^d$ and the state vector X_t lies in $\mathbf{R}^k$ or in a subset of $\mathbf{R}^k$. H and ρ_t are two functions from $\mathbf{R}^d \times \mathbf{R}^k$ to $\mathbf{R}^d$. We assume that the random variables (r.v.) $\theta_0, X_0, X_1, \ldots, X_t, \ldots$ are defined on a probability space $(\Omega, \mathcal{F}, P)$, and we denote the σ-field of events generated by the r.v. $\theta_0, X_0, \ldots, X_t$ by $\mathcal{F}_t$. $(\gamma_t)_{t \in \mathbf{N}}$ is a sequence of non-negative real numbers called the step sizes where $\mathbf{N}$ is the set of natural numbers. The following assumption is made, namely there exists a family $\{\Pi_\theta : \theta \in \mathbf{R}^d\}$ of transition probabilities $\Pi_\theta(x, A)$ on $\mathbf{R}^k$ such that, for any Borel subset A of $\mathbf{R}^k$, we have

$$P[X_{t+1} \in A \mid \mathcal{F}_t] \;=\; \Pi_{\theta_t}(X_t, A)$$

From the above it can be seen that the 2-tuple $(X_t, \theta_t)_{t \geq 0}$ is a Markov process. Its transition probability depends on t (since γ_t and ρ_t depend on t). It is therefore an inhomogeneous Markov process. We prove the convergence of this algorithm under asssumptions which are weaker than in [8].

1.5 Adaptive Control

The issue of adaptive control arises when we don't have knowledge of the underlying transition probabilities or the probability distribution of the immediate cost. In the indirect adaptive control approach we try to estimate the transition probabilities and the expected values of immediate costs, and based on this information we try to choose control stratergies. In direct adaptive control schemes we will be interested in directly finding an optimal control strategy and maybe the optimal cost to go, without estimating the transition probabilities or the expected values of immediate costs. In this dissertation we will be interested in direct adaptive control schemes, in particular we use approximate policy iteration schemes [12, 16] for MDPs. In particular, for the discounted cost problem we will be using temporal difference schemes [16, 19, 20, 26, 50, 54] to estimate the cost to go function and estimate Q-values [16] for further policy improvement. For the average cost problem, we use temporal difference schemes [55] to estimate the differential cost and estimate Q-values for further policy improvement. Q-values are defined in the appropriate chapters for the discounted and average cost problem.

When it comes to adaptive control of POMDPs, the issue becomes even more

complicated. In indirect adaptive control, we should know the cardinality of the underlying state space $\mathcal{S}$ or else it must be estimated. Further the state transition probabilities and observation probabilities along with the expected values of immediate costs need be estimated to arrive at a control strategy. In direct adaptive control of POMDPs we try to arrive at a control law without such estimates. We suggest possible extensions of the direct adaptive control schemes developed for MDPs to the discounted cost POMDP.

Q learning schemes are reinforcement learning schemes based on concepts from value iteration. For Q learning schemes for discounted cost problems see [51, 57]. For Q learning schemes for average cost unichain MDP with a common recurrent state see [1, 17]. For Q learning schemes for SSP see [2, 51]. See [42] for simulation studies of various reinforcement learning schemes for MDPs. For empirical results on average cost reinforcement learning see [38]. For actor-critic reinforcement learning methods for MDPs see [29, 30]. For reinforecement learning schemes for POMDPs see [18, 27, 32, 45]. For an analysis of an adaptive control scheme for a partially observable controlled Markov Chain see [22].

For feature based schemes for large scale dynamic programming see [53]. For real time dynamic programming see [7]. Various learning schemes for solving MDPs are given in [44]. Some interesting algorithms for sequential decision making including solving POMDPs are given in [31]. For linear programming formulations of MDP see [12].

1.6 Organization Of The Dissertation

The rest of the dissertation is organized as follows.

In Chapter 2, we propose various computational schemes for solving POMDPs with finite stage additive cost and infinite horizon discounted cost criteria. Error bounds for the corresponding algorithm are given, and it is further shown that at the expense of additional computational effort the POMDP can be solved as closely to the optimal as desired. We prove that the finite stage optimal costs as well as the optimal cost for the infinite horizon discounted cost problem are both Lipschitz continuous (with domain the belief space, which is the unit simplex of probability distributions over the underlying states) and give bounds for the Lipschitz constant.

In Chapter 3 we prove the convergence of the standard stochastic approximation algorithm presented in [8] under more general assumptions. This in turn can be used to prove the convergence of the TD(λ), the temporal difference schemes discussed later in Chapter 4 and Chapter 5 under more general assumptions.

In Chapter 4 we give an on-line direct adaptive scheme for discounted cost MDP using approximate policy iteration [16] where we use TD(λ) updates to estimate the approximate value function and estimate the corresponding Q-values on-line using a small step stochastic approximation scheme, in order for subsequent policy updating. We use stationary fully randomized policies to approximate deterministic policies, since this allows for exploration and hence lends itself to convergence analysis under weaker assumptions on the transition probabilities. Note that the optimal stationary deterministic policy for sufficiently large (close to 1) discount

factor is a Blackwell optimal policy for the average cost problem [12].

In Chapter 5 we give on-line direct adaptive schemes for average cost unichain MDPs with a common recurrent state using approximate policy iteration. Here also we use temporal difference schemes for estimating the differential cost. Q-value estimates are also obtained on-line using stochastic small step approximation in order for subsequent policy updating.

In Chapter 6 we summarize the contributions of the dissertation and discuss possible extensions of temporal difference schemes to POMDPs.

Appendix A deals with a particular discretization scheme for the unit simplex and provides the bounds on approximation by discretization of the unit simplex. The appendix also deals with some combinatoric results.

Appendix B discusses results on the reachability structure of MDPs.

Appendix C discusses various error bounds for MDPs.

Appendix D discusses temporal difference schemes for SSPs.

Chapter 2

Computational Schemes For Partially Observable Markov Decision Processes With Error Bounds

In this chapter we give computationally feasible techniques for solving the Partially Observable Markov Decision Problem (POMDP) with the infinite horizon total discounted cost criterion. Error bounds for the corresponding algorithm are given, and it is further shown that at the expense of more computational effort the POMDP can be solved as closely to the optimal as desired. The methodology can be easily extended for finite stage additive cost problems with terminal cost. The proofs of all the theorems in this chapter are given in Section 2.5.

2.1 Partially Observable Markov Decision Model

Let $\mathbf{N}$ denote the set of positive integers, $\mathbf{N}_0$ denote the set of non-negative integers. For a set $\mathcal{A}$, $|\mathcal{A}|$ denotes the cardinality of $\mathcal{A}$, whereas for a real number α, $|\alpha|$ denotes the absolute value of α. The homogeneous POMDP [23, 35, 39] can be specified by the tuple $(\mathcal{S}, \mathcal{O}, \mathcal{A}, P, Q, Q_0, p, \mathcal{G})$ where $\mathcal{S} = \{1, ..., n\}$ is the nonempty finite set representing the underlying state space, $\mathcal{O} = \{1, ..., |\mathcal{O}|\}$ is the nonempty finite set of observations and $\mathcal{A} = \{1, ..., |\mathcal{A}|\}$ is the finite nonempty set of actions common to all the states in $\mathcal{S}$. Define $P(j \mid i, a) = p_{ij}(a) \equiv \Pr[s_{t+1} = j \mid s_t = i, u_t = a],\ \forall t \in \mathbf{N}_0,\ i, j \in \mathcal{S}$ and $a \in \mathcal{A}$, where s_t and u_t denote the state and action,

respectively, at time $t \in \mathbf{N}_0$. Here 'Pr' denotes probability. Let $P(a)$ denote the $n \times n$ state transition matrix corresponding to action a with the $(i,j)^{\text{th}}$ entry equal to $p_{ij}(a)$. Let the observation probabilities be given by $Q(l \mid i,a,j) = q(i,a,j,l) \equiv \Pr[y_{t+1} = l \mid s_t = i, u_t = a, s_{t+1} = j],\ \forall t \in \mathbf{N}_0,\ i,j \in \mathcal{S},\ l \in \mathcal{O}$ and $a \in \mathcal{A}$. Here y_{t+1} is the observation made at time $t+1$, after taking action u_t at time t, but before taking action u_{t+1} at time $t+1$. $Q_0(l \mid i) \equiv \Pr[y_0 = l \mid s_0 = i]$ with $i \in \mathcal{S},\ l \in \mathcal{O}$ is the initial observation kernel. Let $p \in \Delta \equiv \{\pi \in \mathbf{R}^n \mid \pi_i \geq 0, \sum_{i=1}^n \pi_i = 1\}$, the $n-1$ dimensional unit simplex in $\mathbf{R}^n$, where π_i is the i^{th} component of the vector $\pi \in \mathbf{R}^n$. Here p represents the initial distribution on the states $\mathcal{S}$ at time $t = 0$. $\mathcal{G}(\cdot \mid i,a,j,l) = \Pr[\cdot \mid s_t = i, u_t = a, s_{t+1} = j, y_{t+1} = l]$ is the probability distribution kernel for the immediate cost $g_t \in \mathbf{R}$ incurred at time t, conditioned on the fact that the state at time t is i, action at time t is a, state at time $t+1$ is j and observation at time $t+1$ is l. For each $i \in \mathcal{S}, a \in \mathcal{A}$, $g(i,a)$ represents the expected immediate cost incurred when action a is taken in state i. $g(i,a,j)$ represents the expected immediate cost incurred at time t given that the current state is $s_t = i$, current action is $u_t = a$ and next state at time $t+1$ is $s_{t+1} = j$. $g(i,a,j)$ is assumed to be finite. Note that $g(i,a) = \sum_{j=1}^n p_{ij}(a) g(i,a,j)$ and that

$$g(i,a,j) = \sum_{l \in \mathcal{O}} q(i,a,j,l) \int_{\mathbf{R}} \omega \mathcal{G}(d\omega \mid i,a,j,l).$$

The POCM (Partially Observable Control Model) evolves as follows. At time $t = 0$, the initial unobservable state s_0 has a prior distribution $p \in \Delta$, and the initial observation y_0 is generated according to the initial observation kernel $Q_0(y_0 \mid s_0)$. If at time $t \in \mathbf{N}_0$, the state is s_t and the control u_t is applied, then an immediate cost

$g_t \in \mathbf{R}$ is incurred and the system moves to the state s_{t+1} according to the transition probability $P(s_{t+1} \mid s_t, u_t)$. The observation $y_{t+1} \in \mathcal{O}$ is generated with probability $Q(y_{t+1} \mid s_t, u_t, s_{t+1})$. A realization of the partially observable system looks like $(s_0, y_0, u_0, g_0, s_1, y_1, u_1, g_1, \ldots) \in \Omega \equiv (\mathcal{S}\mathcal{O}\mathcal{A}\mathbf{R})^\infty$, with s_0 having distribution $p \in \Delta$ and $\{u_t\}$ is a control sequence in $\mathcal{A}$ determined by a control policy. $\mathcal{S}$, $\mathcal{O}$ and $\mathcal{A}$ are endowed with the discrete topology. $\mathbf{R}$ is endowed with the Borel topology. With the metric $\mathsf{d}(\pi, \xi) = \sum_{i=1}^n |\pi_i - \xi_i|$ on Δ (here $\pi, \xi \in \Delta$) the corresponding space (Δ, d) is a Polish space. Note that d is the restriction to Δ of the metric corresponding to the ℓ_1 norm on $\mathbf{R}^n$. Let $h_0 = (p, y_0) \in \mathcal{H}_0$ and the observable history $h_t = (p, y_0, u_0, y_1, u_1, ..., y_{t-1}, u_{t-1}, y_t) \in \mathcal{H}_t$ for $t \in \mathbf{N}$. Here $\mathcal{H}_0 = \Delta\mathcal{O}$ and $\mathcal{H}_t = \mathcal{H}_{t-1}\mathcal{A}\mathcal{O}$ for $t \in \mathbf{N}$, where these spaces are endowed with the corresponding product topologies. An admissible policy for a POMDP is a sequence $\nu = \{\nu_t\}$ such that for each $t \in \mathbf{N}_0$, ν_t is a stochastic kernel on $\mathcal{A}$ given $\mathcal{H}_t$. The set of all admissible policies is denoted by $\mathcal{M}$. In the POCM we assume that the state s_t is not directly observable, to aid us in selecting the action u_t.

A policy $\nu \in \mathcal{M}$ and an initial distribution $p \in \Delta$, together with the stochastic kernels $P, Q, Q_0, \mathcal{G}$, determine a unique probability measure denoted by $\mathcal{P}_p^\nu$ on the space Ω of all possible realizations of the partially observable system [23]. The expectation with respect to this probability is denoted by E_p^ν. The performance criterion for the infinite horizon discounted cost problem is $J(\nu, p) = \mathrm{E}_p^\nu[\sum_{t=0}^\infty \beta^t g_t]$, the expected total discounted cost when the policy $\nu \in \mathcal{M}$ is used and the initial distribution on $\mathcal{S}$ is p. Here $\beta \in [0, 1)$ is the discount factor. The aim of the POMDP is to find a policy $\nu^* \in \mathcal{M}$ such that $J(\nu^*, p) = J^*(p)$, $\forall p \in \Delta$. Here

$J^*(p) = \inf_{\nu \in \mathcal{M}} J(\nu, p),\ p \in \Delta$, is the optimal cost function. For a finite stage problem with k stages the objective function is $J_k(\nu, p) = \mathrm{E}_p^{\nu}[\sum_{t=0}^{k-1} \beta^t g_t + \beta^k r(s_k)]$, where $\beta^k r(i),\ i \in \mathcal{S}$, is the terminal cost of being in state i at the k^{th} instant. In the finite stage problem the restriction that $\beta < 1$ can be removed, i.e. $\beta \in [0, \infty)$. In the finite horizon problem with k stages, as far as the policy is concerned, only $\{\nu_0, \nu_1, \ldots, \nu_{k-1}\}$ is of interest. Let the optimal k stage cost function be defined as $J_k^*(p) = \inf_{\nu \in \mathcal{M}} J_k(\nu, p)$.

2.2 Equivalent Fully Observable MDP

It is well known [5, 23, 47, 49] that the useful information in h_t can be encapsulated in a vector $p_t \in \Delta$ for determining the best action u_t at time t, $\forall t \in \mathbf{N}_0$, (i.e. p_t is a sufficient statistic), and the POMDP can be recast into an equivalent completely observed Markov Decision Process (MDP) with stationary structure [10, 11] having as its state space the uncountable set Δ. Here $[p_t]_i = \Pr[s_t = i \mid h_t]$ for each $i \in \mathcal{S}$. This may be computed recursively as follows (here $h_0 = (p, y_0)$):

$$[p_0]_i = \frac{Q_0(y_0|i)\,[p]_i}{\sum_{j=1}^{n} [p]_j\, Q_0(y_0|j)} \qquad \text{for} \quad i \in \{1, ..., n\}.$$

For $t \in \mathbf{N}_0$ the following Bayesian update rule is used :

$$\begin{aligned} p_{t+1}^T &= F(p_t, u_t, y_{t+1}) \qquad \text{where} \\ F(\pi, a, o) &= \frac{\pi^T \bar{P}^o(a)}{\sigma(\pi, a, o)}; \qquad \pi \in \Delta, a \in \mathcal{A}, o \in \mathcal{O}. \end{aligned}$$

Here $\bar{P}^o(a) \equiv P(a) \odot \bar{Q}_a^o$ where the operator $\odot$ denotes term by term multiplication; i.e. $[\bar{P}^o(a)]_{i,j} = [P(a)]_{i,j}\,[\bar{Q}_a^o]_{i,j}$. Also $\sigma(\pi, a, o) = \pi^T \bar{P}^o(a)\,\underline{1}$. The superscript of

π^T denotes transposition of the vector π and $\bar{Q}_a^l$ is the $n \times n$ matrix with $(i,j)^{\text{th}}$ entry equal to $q(i,a,j,l)$. $\underline{1} \in \mathbf{R}^n$ is the vector with all components equal to 1. $\sigma(\pi, a, o)$ is the probability of observing $o \in \mathcal{O}$ at time $t+1$, given prior distribution $\pi \in \Delta$ on $\mathcal{S}$ at time t and that action $a \in \mathcal{A}$ is taken at time t, for any $t \in \mathbf{N}_0$. $[F(\pi, a, o)]^T$ denotes the aposterior probability on the states $\mathcal{S}$ at time $t+1$, given prior probability $\pi \in \Delta$ on $\mathcal{S}$ at time t, action $a \in \mathcal{A}$ is executed at time t and observation $o \in \mathcal{O}$ is made at time $t+1$. The above relations on the Bayesian transition function F and the observation probability σ, may be arrived at as follows. Notice that

$$\Pr[s_{t+1} \mid p_t, u_t, y_{t+1}] = \frac{\Pr[s_{t+1}, y_{t+1} \mid p_t, u_t]}{\Pr[y_{t+1} \mid p_t, u_t]}.$$

For $j \in \mathcal{S}$, $o \in \mathcal{O}$, $a \in \mathcal{A}$, we may compute

$$\begin{aligned}
&\Pr[s_{t+1} = j, y_{t+1} = o \mid p_t = \pi, u_t = a] \\
&= \sum_{i=1}^{n} \Pr[s_{t+1} = j, y_{t+1} = o \mid s_t = i, u_t = a]\, \pi_i \\
&= \sum_{i=1}^{n} \Pr[y_{t+1} = o \mid s_{t+1} = j, s_t = i, u_t = a] \Pr[s_{t+1} = j \mid s_t = i, u_t = a]\, \pi_i \\
&= \sum_{i=1}^{n} \Pr[y_{t+1} = o \mid s_{t+1} = j, s_t = i, u_t = a]\, p_{ij}(a)\, \pi_i \\
&= \sum_{i=1}^{n} q(i,a,j,o)\, p_{ij}(a)\, \pi_i.
\end{aligned}$$

Also

$$\begin{aligned}
\Pr[y_{t+1} = o \mid p_t = \pi, u_t = a] &= \sum_{j=1}^{n} \Pr[s_{t+1} = j, y_{t+1} = o \mid p_t, u_t = a] \\
&= \sum_{j=1}^{n} \sum_{i=1}^{n} q(i,a,j,o)\, p_{ij}(a)\, \pi_i.
\end{aligned}$$

With

$$\bar{P}^o(a) = P(a) \odot \bar{Q}_a^o$$

we obtain the desired result.

The transition kernel for the equivalent MDP [5, 23] with state space Δ is given by :

$$\mathcal{K}(D \mid \pi, a) = \sum_{o \in \mathcal{O}} \sigma(\pi, a, o)\, \boldsymbol{\mathcal{I}}_{[[F(\pi,a,o)]^T \in D]} \tag{2.1}$$

with $D \in \mathcal{B}(\Delta)$, where $\mathcal{B}(\Delta)$ is the Borel sigma field on Δ. Here $\boldsymbol{\mathcal{I}}$ is the indicator function. In fact the above equation 2.1 holds for any $D \subset \Delta$. We could have chosen as our σ-field the collection of arbitrary subsets of Δ. The expected immediate cost for taking action $a \in \mathcal{A}$ from state $\pi \in \Delta$ for this MDP is given by $\pi^T \bar{g}(a)$, where $\bar{g}(a) = (g(1,a), ..., g(n,a))^T$. The original discounted cost criterion POMDP can be solved by solving this new MDP with the infinite horizon discounted cost criterion (using the same discount factor β) [35, 39]. For the k stage problem the terminal cost at time k for this new MDP at state $\pi \in \Delta$ is set to be $\beta^k (\pi^T r)$, where $r = (r(1), ..., r(n))^T \in \mathbf{R}^n$.

2.3 Lipschitz Continuity Of Value Functions

The optimal value function for the infinite horizon discounted cost problem on this equivalent MDP, denoted by $V^* : \Delta \to \mathbf{R}$, is known to be concave and continuous [6, 23, 33, 36]. Also, the existence of a stationary non-randomized optimal Markov policy for this equivalent MDP is guaranteed. In fact

$$J^*(p) \;=\; \sum_{o \in \mathcal{O}} \left[\left(\sum_{j \in \mathcal{S}} Q_0(o|j)\, [p]_j \right) V^*(\varpi(p,o)) \right]$$

where

$$[\varpi(p,o)]_i \;=\; \frac{Q_0(o|i)\,[p]_i}{\sum_{j\in\mathcal{S}} Q_0(o|j)\,[p]_j}. \tag{2.2}$$

The function J^* is also continuous and concave on Δ. Let $B(\Delta)$ be the set of all bounded real valued functions on Δ with the distance between $U, V \in B(\Delta)$ given by

$$\boldsymbol{\rho}(U,V) = \sup_{\pi\in\Delta} |U(\pi) - V(\pi)|.$$

With this metric $B(\Delta)$ is a complete metric space.

Define the function $\mathsf{h} : \Delta \times \mathcal{A} \times B(\Delta) \to \mathbf{R}$ by

$$\mathsf{h}(\pi,a,V) = \pi^T \bar{g}(a) + \beta \sum_{o\in\mathcal{O}} \sigma(\pi,a,o)\, V([F(\pi,a,o)]^T)$$

where $\pi \in \Delta,\ a \in \mathcal{A},\ V \in B(\Delta)$. Let the function $H : B(\Delta) \to B(\Delta)$ be defined by

$$(H\,V)\ (\pi) = \min_{a\in\mathcal{A}} \mathsf{h}(\pi,a,V)$$

where $\pi \in \Delta$ and $V \in B(\Delta)$. For a control function $\delta : \Delta \to \mathcal{A}$, define the corresponding mapping $H_\delta : B(\Delta) \to B(\Delta)$ by

$$(H_\delta\,V)\ (\pi) = \mathsf{h}(\pi,\delta(\pi),V)$$

where $\pi \in \Delta$ and $V \in B(\Delta)$. For $U, V \in B(\Delta)$ we denote $U \le V$ if and only if $U(\pi) \le V(\pi),\ \forall\pi \in \Delta$. Note that H_δ and H are monotone operators, i.e. $U, V \in B(\Delta)$, $U \le V$ implies $H_\delta U \le H_\delta V$ and $HU \le HV$. Also if $\beta \in [0,1)$, H_δ and H are contraction mappings with contraction coefficients β; i.e. for $U, V \in B(\Delta),\ \boldsymbol{\rho}(HU,HV) \le \beta\boldsymbol{\rho}(U,V)$ and $\boldsymbol{\rho}(H_\delta U, H_\delta V) \le \beta\boldsymbol{\rho}(U,V)$.

A control function $\delta : \Delta \to \mathcal{A}$ is said to be "greedy" for a $V \in B(\Delta)$ if

$H_\delta V = HV$, that is

$$\delta(\pi) = \arg\min_{a \in \mathcal{A}} \mathsf{h}(\pi, a, V) \qquad \forall \pi \in \Delta.$$

(If there are multiple minimizing arguments, we could pick any of them.)

Also if $\beta \in [0, 1)$, H has unique fixed point V^*, the optimal value function under the infinite horizon discounted cost criterion for the equivalent MDP with state space Δ. Similarly for a stationary policy δ (using control function δ) H_δ has unique fixed point V^δ, the value function corresponding to policy δ for the discounted cost problem defined on the equivalent MDP with state space Δ. Note that a stationary policy δ^* such that $H_{\delta^*} V^* = HV^*$ is optimal for the equivalent MDP. Define $\Gamma_0 \equiv \{r\}$, the singleton set with its element in $\mathbf{R}^n$.

Let $V_k^* : \Delta \to \mathbf{R},\ k = 0, 1, 2, ...$ denote the optimal value function for the k stage problem. For the finite stage problem, $V_0^*(\pi) = \pi^T r$ and $V_k^*,\ k = 1, 2, ...$ can be computed in that order by value iteration [23, 36], namely $V_k^* = H\, V_{k-1}^*$. The functions $\{V_k^*\}$ are known to be piecewise linear and concave and can each be represented as the minimum of a finite number of linear functions [35, 39, 48], i.e.

$$V_k^*(\pi) = \min_{W \in \Gamma_k} \pi^T W$$

$k \in \mathbf{N}_0$, $\pi \in \Delta$. For each $k \in \mathbf{N}_0$, Γ_k is a finite set of vectors in $\mathbf{R}^n$ and for $k \in \mathbf{N}$, each $W \in \Gamma_k$ has the form $W = \bar{g}(a) + \beta \sum_{o \in \mathcal{O}} \bar{P}^o(a) W_{\varphi_o}$; for some $a \in \mathcal{A}$. Here φ_o is some indexing into the set Γ_{k-1} so that $W_{\varphi_o} \in \Gamma_{k-1}$. But the number of linear functions needed to represent V_k^* or equivalently $|\Gamma_k|$ may grow exponentially fast (at most $|\mathcal{A}|^{\frac{|\mathcal{O}|^k - 1}{|\mathcal{O}| - 1}}$ for V_k^*); to get a minimal representation for the exact values of V_k^* one may have to use linear programming as in Sondik's method [35, 39, 48] or

the more recent method due to Littman [28, 31], and this may be computationally expensive. Note that

$$J_k^*(p) = \sum_{o \in \mathcal{O}} \left[(\sum_{j \in \mathcal{S}} Q_0(o \mid j)[p]_j)\, V_k^*(\varpi(p,o)) \right]$$

where $\varpi(p,o)$ is defined as in equation 2.2. J_k^* is also piecewise linear and concave and can be represented as the minimum of a finite number of linear functions. When $\beta \in [0,1)$, V_k^* converges to V^* uniformly as $k \to \infty$ irrespective of the terminal cost which is used. However we may assume that the terminal cost is zero when we use value iteration to approximate V^*.

Define the following constants :

$$G_{\max} = \max_{i \in \mathcal{S}} \max_{a \in \mathcal{A}} g(i,a), \quad G_{\min} = \min_{i \in \mathcal{S}} \min_{a \in \mathcal{A}} g(i,a),$$

$$r_{\max} = \max_{i \in \mathcal{S}} r(i), \qquad r_{\min} = \min_{i \in \mathcal{S}} r(i),$$

$$C = G_{\max} - G_{\min}.$$

Fix integer $k > 1$. Let control functions δ_t^* be such that $H_{\delta_t^*} V_t^* = H V_t^*$ for $t = 0, 1, \ldots, k-1$. Then the Markov policy $\{\delta_{k-1}^*, \delta_{k-2}^*, \ldots, \delta_1^*, \delta_0^*\}$ is optimal for the k stage problem of the equivalent MDP. Here control function δ_t^* is used at stage $(k-1-t)$ for $k = 0, 1, \ldots, k-1$.

Theorem 2.1 *$\{V_k^*\}$ and V^* are Lipschitz continuous, and a Lipschitz constant for V^* is $\frac{C}{2}\frac{1}{1-\beta}$. In the k stage finite horizon case with non-zero terminal cost, a Lipschitz constant for V_k^* is given by $\frac{C}{2}(\sum_{t=0}^{k-1} \beta^t) + \frac{1}{2}\beta^k(r_{\max} - r_{\min})$. Note that C is a constant, independent of P and Q, that depends only on the expected immediate cost.* □

Note that $\sum_{t=0}^{k-1} \beta^t = \frac{1-\beta^k}{1-\beta}$ when $\beta \neq 1$, and $\sum_{t=0}^{k-1} \beta^t = k$ for $\beta = 1$.

2.4 Approximation By Discretization

A method for approximating V^* is given by finding the value function for a finite state MDP derived from the uncountable state MDP by partitioning the state space Δ in the spirit of [24]. However the theorems given in [24] are not directly applicable to this problem since some of the assumptions are not satisfied by the present problem, for example the Bayesian transition function is not Lipschitz continuous in general. But we use the Lipschitz continuity of the optimal value functions $\{V_k^*\}$ and V^* to circumvent this. Let $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_m\}$ be a finite partition of Δ, where $\mathcal{D}_i$, $i = 1, 2..., m$, are disjoint measurable subsets of Δ such that $\Delta = \bigcup_{i=1}^m \mathcal{D}_i$. For each $i = 1, ..., m$, let $d_i \in \mathcal{D}_i$ be an arbitrary representative point in $\mathcal{D}_i$. A new finite state MDP is constructed with the states being the points in the grid $\mathcal{E} = \{d_i \mid i = 1, ..., m\}$, the transition probabilities being $p_{ij}^{\mathcal{D}}(a) = \mathcal{K}(\mathcal{D}_j \mid d_i, a),\ i, j \in \{1, ..., m\},\ a \in \mathcal{A}$, with the stochastic kernel $\mathcal{K}$ as defined earlier in equation 2.1. The immediate cost function is given by $g^{\mathcal{D}}(i, a) = d_i^T \bar{g}(a)$ for $i \in \{1, ..., m\},\ a \in \mathcal{A}$. Let $\bar{V}^{\mathcal{D}} \in \mathbf{R}^m$ be the optimal value function for this infinite horizon discounted cost minimization problem with the same $\beta \in [0, 1)$. For the finite horizon problem we may assign a terminal cost $r^{\mathcal{D}}(i) = d_i^T r;\ i \in \{1, ..., m\}$. Let $\bar{V}_k^{\mathcal{D}} \in \mathbf{R}^m;\ k = 0, 1, ...,$ denote the finite k stage optimal costs obtained by value iteration, i.e.,

$$[\bar{V}_0^{\mathcal{D}}]_i = r^{\mathcal{D}}(i) \qquad \forall i \in \{1, ..., m\}.$$

For $k \geq 1$,

$$[\bar{V}_k^{\mathcal{D}}]_i \;=\; \min_{a \in \mathcal{A}} \left\{ g^{\mathcal{D}}(i,a) + \beta \sum_{j=1}^{m} p_{ij}^{\mathcal{D}}(a) [\bar{V}_{k-1}^{\mathcal{D}}]_j \right\} \qquad \forall i \in \{1, ..., m\}.$$

Note that $\bar{V}_k^{\mathcal{D}} \underset{k \to \infty}{\longrightarrow} \bar{V}^{\mathcal{D}}$ and

$$\| \bar{V}^{\mathcal{D}} - \bar{V}_k^{\mathcal{D}} \| \leq \; \beta^k \; \| \bar{V}^{\mathcal{D}} - \bar{V}_0^{\mathcal{D}} \|$$

where $\| \cdot \|$ denotes the sup-norm given by $\| \bar{V} \| = \max_{i \in \{1,2,...,m\}} |[\bar{V}]_i|$ for $\bar{V} \in \mathbf{R}^m$. Extend $\bar{V}^{\mathcal{D}}$ to the whole of Δ by taking $V^{\mathcal{D}}(\pi) = [\bar{V}^{\mathcal{D}}]_i$ if $\pi \in \mathcal{D}_i$. A similar piecewise constant extension can be performed to obtain $V_k^{\mathcal{D}}(\pi) = [\bar{V}_k^{\mathcal{D}}]_i$ if $\pi \in \mathcal{D}_i$, for the finite stage problem. Define the diameter of the partition $\mathcal{D}$ by $\mathrm{Diam}(\mathcal{D}) \equiv \max_{1 \leq i \leq m} \sup_{\pi, \xi \in \mathcal{D}_i} \mathsf{d}(\pi, \xi)$.

Theorem 2.2 *For the infinite horizon discounted cost problem,*

$$\boldsymbol{\rho}(V^{\mathcal{D}}, V^*) \leq \frac{C}{2} \frac{\mathrm{Diam}(\mathcal{D})}{(1-\beta)^2}.$$

For the finite k stage problem,

$$\boldsymbol{\rho}(V_k^*, V_k^{\mathcal{D}}) \leq \left[\frac{C}{2} (\sum_{t=0}^{k-1} (t+1)\beta^t) + \frac{(k+1)}{2} \beta^k (r_{\max} - r_{\min}) \right] \mathrm{Diam}(\mathcal{D}).$$

□

Note that

$$\begin{aligned} \sum_{t=0}^{k-1} (t+1)\beta^t \;&=\; \frac{d}{d\beta} \left(\sum_{t=0}^{k} \beta^t \right) \;=\; \frac{d}{d\beta} \left(\frac{1-\beta^{k+1}}{1-\beta} \right) \\ &=\; \frac{1 - (k+1)\beta^k + k\beta^{k+1}}{(1-\beta)^2}. \end{aligned}$$

Also for $\beta = 1$,

$$\sum_{t=0}^{k-1}(t+1)\beta^t = \frac{k(k+1)}{2}.$$

Now $\bar{V}^{\mathcal{D}} \in \mathbf{R}^m$ may be solved by any of the standard methods like policy iteration or may be approximated as closely as desired by value iteration over a finite, though large, number of steps [10, 12]. The following result (see Lemma C.2 in Appendix C) which is an extension of the results in [16, 46] may be used to find a suboptimal stationary nonrandomized policy for the infinite horizon discounted cost problem with state space Δ.

Lemma 2.1 *Let $U \in B(\Delta)$ be such that $\boldsymbol{\rho}(U, V^*) \leq \epsilon$. Assume that $V^\delta : \Delta \to \mathbf{R}$ is the value function for the infinite horizon discounted cost problem (with state space Δ) obtained by following the stationary non-randomized Markov policy δ, where $\delta : \Delta \to \mathcal{A}$ corresponds to the one-step "near greedy" control function obtained while doing approximate dynamic programming update [12] on U (i.e. $\boldsymbol{\rho}(H_\delta U, HU) \leq \varepsilon$). Then $\boldsymbol{\rho}(V^*, V^\delta) \leq \frac{2\epsilon\beta+\varepsilon}{1-\beta}$.*

□

(With slight abuse of notation we use δ to represent both the control function $\delta : \Delta \to \mathcal{A}$ as well as the stationary policy). Lemma 2.1 along with Theorem 2.2 (which gives the bound for $\boldsymbol{\rho}(V^{\mathcal{D}}, V^*)$) can be used to find a stationary non-randomized suboptimal policy (which can be made as close to the optimal as desired) for the MDP with state space Δ. Similar bounds for the approximate value functions for finite stage problems, along with suboptimal nonrandomized Markov policies (though not guaranteed to be stationary), may be obtained.

For $k \in \mathbf{N}_0$, define the control functions $\delta_k : \Delta \to \mathcal{A}$ by

$$\delta_k(\pi) = \arg\min_{a \in \mathcal{A}} \mathsf{h}(\pi, a, V_k^{\mathcal{D}}) \qquad \forall \pi \in \Delta.$$

Let $\Psi_k \equiv \{\delta_{k-1}, \delta_{k-2}, \ldots, \delta_0\}$, denote a Markov policy [35] for the k stage equivalent MDP with state space Δ. Under this policy, for a k stage problem, the control function δ_{k-1-t} is used to choose the control action at the t^{th} stage for $t \in \{0, 1, \ldots, k-1\}$. Let $V_k^{\Psi} : \Delta \to \mathbf{R}$ denote the corresponding value function for the k stage problem while using the control policy Ψ_k, with $V_0^{\Psi} = V_0^*$. It is easy to see that for $k \in \mathbf{N}$, $V_k^{\Psi} = H_{\delta_{k-1}} V_{k-1}^{\Psi}$. The following result holds.

Theorem 2.3 *For $k \in \mathbf{N}$, the k stage value function V_k^{Ψ} corresponding to the policy Ψ_k satisfies the relationship*

$$\boldsymbol{\rho}(V_k^{\Psi}, V_k^*) \leq \left[\frac{C}{2} (\sum_{t=0}^{k-1} t(t+1)\beta^t) \; + \; \frac{k(k+1)}{2} \beta^k (r_{\max} - r_{\min}) \right] \text{Diam}(\mathcal{D}).$$

□

We also give another Markov policy defined as follows. For $k \in \mathbf{N}_0$, define the control functions $\hat{\delta}_k : \Delta \to \mathcal{A}$ by,

$$\hat{\delta}_k(\pi) = \arg\min_{a \in \mathcal{A}} \mathsf{h}(d_i, a, V_k^{\mathcal{D}}) \qquad \forall \pi \in \mathcal{D}_i$$

for $i \in \{1, \ldots, m\}$. Observe that $\mathsf{h}(d_i, a, V_k^{\mathcal{D}}) = g^{\mathcal{D}}(i, a) + \beta \sum_{j=1}^{m} p_{ij}^{\mathcal{D}}(a) [\bar{V}_k^{\mathcal{D}}]_j$. Let $\hat{\Psi}_k \equiv \{\hat{\delta}_{k-1}, \hat{\delta}_{k-2}, \ldots, \hat{\delta}_0\}$, denote another Markov policy for the k stage equivalent MDP with state space Δ. Under this policy, for a k stage problem, the control function $\hat{\delta}_{k-1-t}$ is used to choose the control action at the t^{th} stage for $t \in \{0, 1, \ldots, k-1\}$. Let $V_k^{\hat{\Psi}} : \Delta \to \mathbf{R}$ denote the value function for the k stage problem while using the control policy $\hat{\Psi}_k$, with $V_0^{\hat{\Psi}} = V_0^*$. It may be seen that for $k \in \mathbf{N}$, $V_k^{\hat{\Psi}} = H_{\hat{\delta}_{k-1}} V_{k-1}^{\hat{\Psi}}$.

Theorem 2.4 *For $k \in \mathbf{N}$ the k stage value function $V_k^{\hat{\Psi}}$ corresponding to the policy $\hat{\Psi}_k$ satisfies the relationship*

$$\boldsymbol{\rho}(V_k^{\hat{\Psi}}, V_k^*) \leq \left[\frac{C}{2}(\sum_{t=0}^{k-1}(t+1)(t+2)\beta^t) \; + \; \frac{k(k+3)}{2}\beta^k(r_{\max} - r_{\min})\right] \text{Diam}(\mathcal{D}).$$

□

Note that

$$\begin{aligned}
\sum_{t=0}^{k-1}(t+1)(t+2)\beta^t &= \frac{d}{d\beta}\left(\sum_{t=0}^{k}(t+1)\beta^t\right) \\
&= \frac{d}{d\beta}\left(\frac{1-(k+2)\beta^{k+1}+(k+1)\beta^{k+2}}{(1-\beta)^2}\right) \\
&= \frac{2-(k+1)(k+2)\beta^k+2k(k+2)\beta^{k+1}-k(k+1)\beta^{k+2}}{(1-\beta)^3}.
\end{aligned}$$

Also for $\beta = 1$,

$$\sum_{t=0}^{k-1}(t+1)(t+2)\beta^t = \frac{k(k+1)(k+2)}{3}.$$

Let the control function $\hat{\delta} : \Delta \to \mathcal{A}$ be defined by,

$$\hat{\delta}(\pi) = \arg\min_{a\in\mathcal{A}} \mathsf{h}(d_i, a, V^{\mathcal{D}}), \qquad \forall \pi \in \mathcal{D}_i$$

for $i \in \{1, \ldots, m\}$. Note that $\mathsf{h}(d_i, a, V^{\mathcal{D}}) = g^{\mathcal{D}}(i,a) + \beta \sum_{j=1}^m p_{ij}^{\mathcal{D}}(a)[\bar{V}^{\mathcal{D}}]_j$. This control function is essentially the "piecewise constant" extension to Δ of the optimal stationary control function of the discretized finite state MDP. Let $V^{\hat{\delta}} : \Delta \to \mathbf{R}$, denote the value function obtained for the equivalent MDP under the infinite horizon discounted cost criterion while following the stationary "policy" $\hat{\delta}$. When $\beta \in [0, 1)$ the following corollary to Theorem 2.4 may be obtained. The proof can be adapted from that of Theorem 2.4. We omit the details.

Corollary 2.1 *The value function $V^{\hat{\delta}}$ for the infinite horizon dicounted cost problem, obtained while following the stationary "policy" $\hat{\delta}$, satisfies the relationship*

$$\boldsymbol{\rho}(V^{\hat{\delta}}, V^*) \leq \frac{C}{(1-\beta)^3} \operatorname{Diam}(\mathcal{D}).$$

□

We mention in passing that our analysis can be used for finding analytical error bounds for the schemes discussed in [36] and may be used to show that by making the grid finer in [36] we can obtain suboptimal policies which are as close to the optimal as desired.

2.5 Proof Of Theorems

In this section we give the proofs of Theorem 2.1 and Theorem 2.2, and outline the proofs of Theorem 2.3 and Theorem 2.4.

Definition 2.1 *For $W \in \mathbf{R}^n$ define*

$$\operatorname{slope}(W) \equiv \max_{i \in \{1,\cdots,n\}} [W]_i - \min_{i \in \{1,\cdots,n\}} [W]_i.$$

□

Here $[W]_i$ is the i^{th} component of W. Note that the following three properties of slope follow easily from the definition.

1. For any $n \times n$ stochastic matrix $\tilde{P}$ and any $W \in \mathbf{R}^n$, $\operatorname{slope}(\tilde{P}\, W) \leq \operatorname{slope}(W)$.

2. For any $\lambda \in \mathbf{R}$ and $W \in \mathbf{R}^n$, $\text{slope}(\lambda W) = |\lambda| \text{ slope}(W)$.

3. For any $W, \tilde{W} \in \mathbf{R}^n$, $\text{slope}(W + \tilde{W}) \leq \text{slope}(W) + \text{slope}(\tilde{W})$.

Lemma 2.2 *For any given $\pi, \xi \in \Delta$ and $W \in \mathbf{R}^n$*

$$|\pi^T W - \xi^T W| \leq \frac{1}{2} \mathsf{d}(\pi, \xi) \text{ slope}(W).$$

□

Proof of Lemma 2.2

Let $W_{\max} \equiv \max_{i \in \{1,\cdots,n\}} [W]_i$, and $W_{\min} \equiv \min_{i \in \{1,\cdots,n\}} [W]_i$. Let

$$\begin{aligned} \bar{\mathcal{I}} &= \{i \in \{1, \cdots, n\} : \pi_i > \xi_i\}, \\ \bar{\mathcal{J}} &= \{i \in \{1, \cdots, n\} : \pi_i < \xi_i\}. \end{aligned}$$

Note that

$$\sum_{i \in \bar{\mathcal{I}}} (\pi_i - \xi_i) + \sum_{i \in \bar{\mathcal{J}}} (\pi_i - \xi_i) = \sum_{i \in \{1,\cdots n\}} (\pi_i - \xi_i) = 0.$$

This implies that

$$\sum_{i \in \bar{\mathcal{I}}} (\pi_i - \xi_i) = \sum_{i \in \bar{\mathcal{J}}} (\xi_i - \pi_i) = \frac{1}{2} \mathsf{d}(\pi, \xi).$$

Now

$$\sum_{i \in \bar{\mathcal{I}}} (\pi_i - \xi_i)[W]_i \in \left[\frac{1}{2} \mathsf{d}(\pi, \xi) W_{\min}, \frac{1}{2} \mathsf{d}(\pi, \xi) W_{\max} \right]$$

and

$$\sum_{j \in \bar{\mathcal{J}}} (\xi_j - \pi_j)[W]_j \in \left[\frac{1}{2} \mathsf{d}(\pi, \xi) W_{\min}, \frac{1}{2} \mathsf{d}(\pi, \xi) W_{\max} \right].$$

Hence

$$
\begin{aligned}
&\left|\pi^T W - \xi^T W\right| \\
&= \left|\sum_{i\in\bar{\mathcal{I}}}(\pi_i - \xi_i)[W]_i - \sum_{j\in\bar{\mathcal{J}}}(\xi_j - \pi_j)[W]_j\right| \\
&\leq \frac{1}{2}\mathsf{d}(\pi,\xi)(W_{\max} - W_{\min}) \\
&= \frac{1}{2}\,\mathsf{d}(\pi,\xi)\ \mathrm{slope}(W).
\end{aligned}
$$

□

Let Γ be a finite nonempty set of vectors in $\mathbf{R}^n$. Define

$$
\begin{aligned}
\max_\Gamma &\equiv \max_{W\in\Gamma}\max_{i\in\{1,\cdots,n\}}[W]_i, \\
\min_\Gamma &\equiv \min_{W\in\Gamma}\min_{i\in\{1,\cdots,n\}}[W]_i.
\end{aligned}
$$

Lemma 2.3 *For any $a \in \mathcal{A}$ consider the vector $[\bar{g}(a) + \beta\sum_{o\in\mathcal{O}}\bar{P}^o(a)W_{\varphi_o}] \in \mathbf{R}^n$, where φ_o is an indexing into the set Γ dependent on o so that $W_{\varphi_o} \in \Gamma$. Then for $i \in \{1,\cdots,n\}$*

$$
\min_{j\in\mathcal{S}} g(j,a) + \beta\min_\Gamma \leq \left[\bar{g}(a) + \beta\sum_{o\in\mathcal{O}}\bar{P}^o(a)W_{\varphi_o}\right]_i
$$

and

$$
\max_{j\in\mathcal{S}} g(j,a) + \beta\max_\Gamma \geq \left[\bar{g}(a) + \beta\sum_{o\in\mathcal{O}}\bar{P}^o(a)W_{\varphi_o}\right]_i.
$$

□

Proof of Lemma 2.3

Observe that

$$
\begin{aligned}
\left[\sum_{o\in\mathcal{O}}\bar{P}^o(a)W_{\varphi_o}\right]_i &= \sum_{o\in\mathcal{O}}\sum_{j=1}^n[\bar{P}^o(a)]_{i,j}[W_{\varphi_o}]_j \\
&= \sum_{o\in\mathcal{O}}\sum_{j=1}^n p_{ij}(a)q(i,a,j,o)[W_{\varphi_o}]_j \\
&= \sum_{j=1}^n p_{ij}(a)\sum_{o\in\mathcal{O}} q(i,a,j,o)[W_{\varphi_o}]_j
\end{aligned}
$$

Also $\min_\Gamma \leq \sum_{o\in\mathcal{O}} q(i,a,j,o)[W_{\varphi_o}]_j \leq \max_\Gamma$ for each $i,j \in \{1,\ldots,n\}$. Hence it follows that for each $i \in \{1,\ldots,n\}$

$$\min_\Gamma \leq \left[\sum_{o\in\mathcal{O}} \bar{P}^o(a) W_{\varphi_o}\right]_i \leq \max_\Gamma.$$

Now it may be seen that for $i \in \{1,\cdots,n\}$

$$\min_{j\in\mathcal{S}} g(j,a) + \beta \min_\Gamma \leq \left[\bar{g}(a) + \beta \sum_{o\in\mathcal{O}} \bar{P}^o(a) W_{\varphi_o}\right]_i$$

and

$$\max_{j\in\mathcal{S}} g(j,a) + \beta \max_\Gamma \geq \left[\bar{g}(a) + \beta \sum_{o\in\mathcal{O}} \bar{P}^o(a) W_{\varphi_o}\right]_i.$$

It may also be seen that

$$\begin{aligned}\text{slope}\left(\bar{g}(a) + \beta \sum_{o\in\mathcal{O}} \bar{P}^o(a) W_{\varphi_o}\right) \leq \\ \left(\max_{j\in\mathcal{S}} g(j,a) - \min_{j\in\mathcal{S}} g(j,a)\right) + \beta\,(\max_\Gamma - \min_\Gamma).\end{aligned} \tag{2.3}$$

□

Corollary 2.2 *The function from* $\Delta \to \mathbf{R}$ *defined by*

$$\pi \mapsto \pi^T[\bar{g}(a) + \beta \sum_{o\in\mathcal{O}} \bar{P}^o(a) W_{\varphi_o}] = \pi^T \bar{g}(a) + \beta \sum_{o\in\mathcal{O}} \sigma(\pi,a,o)\, F(\pi,a,o)\, W_{\varphi_o}$$

is Lipschitz continuous with a Lipschitz constant $\frac{1}{2}[(\max_{j\in\mathcal{S}} g(j,a) - \min_{j\in\mathcal{S}} g(j,a)) + \beta\,(\max_\Gamma - \min_\Gamma)]$.

□

Proof of Corollary 2.2

This follows from Lemma 2.3 and Lemma 2.2.

□

Let $\min_{\Gamma_k} \equiv \min_{W \in \Gamma_k} \min_{i \in \{1,\cdots,n\}} [W]_i$ and $\max_{\Gamma_k} \equiv \max_{W \in \Gamma_k} \max_{i \in \{1,\cdots,n\}} [W]_i$, where Γ_k was defined earlier. Note that $\Gamma_0 = \{r\}$, and hence $\min_{\Gamma_0} = r_{\min}$ and $\max_{\Gamma_0} = r_{\max}$.

Proof of Theorem 2.1

For $k \geq 1$, if $W \in \Gamma_k$ then for some $a \in \mathcal{A}$, $W = \bar{g}(a) + \beta \sum_{o \in \mathcal{O}} \bar{P}^o(a) W_{\varphi_o}$, where φ_o is an indexing into the set Γ_{k-1} dependent on o so that $W_{\varphi_o} \in \Gamma_{k-1}$. This together with Lemma 2.3 implies

$$G_{\min} + \beta \min_{\Gamma_{k-1}} \leq \min_{\Gamma_k} \leq \max_{\Gamma_k} \leq G_{\max} + \beta \max_{\Gamma_{k-1}}.$$

By induction it may be seen that

$$G_{\min} \sum_{t=0}^{k-1} \beta^t \; + \; \beta^k \min_{\Gamma_0} \;\; \leq \;\; \min_{\Gamma_k}$$

and

$$G_{\max} \sum_{t=0}^{k-1} \beta^t \; + \; \beta^k \max_{\Gamma_0} \;\; \geq \;\; \max_{\Gamma_k}.$$

This in turn implies that for $k \geq 1$, if $W \in \Gamma_k$ then $\text{slope}(W) \leq C(\sum_{t=0}^{k-1} \beta^t) + \beta^k(\max_{\Gamma_0} - \min_{\Gamma_0})$. Now $V_k^* = \min_{W \in \Gamma_k} \pi^T W$. Since V_k^* is the minimum of a finite number of Lipschitz continuous functions defined on the convex subset Δ of $\mathbf{R}^n$, V_k^* itself is Lipschitz continuous with a Lipschitz constant which is the largest among the constituent ones.

Hence by Lemma 2.2, V_k^* is Lipschitz continuous with a Lipschitz constant $\frac{C}{2}(\sum_{t=0}^{k-1}\beta^t) + \frac{\beta^k}{2}(\max_{\Gamma_0} - \min_{\Gamma_0})$. When $\beta \in [0,1)$, V_k^* converges to V^* uniformly on Δ at a geometric rate governed by β. Hence taking the limit gives that V^* is Lipschitz continuous with Lipschitz constant $\frac{C}{2}\frac{1}{1-\beta}$.

□

Proof of Theorem 2.2

From Theorem 2.1, Lemma 2.2 and the definition of $V_0^{\mathcal{D}}$ it follows that $\pi \in \mathcal{D}_i,\ i \in \{1,\cdots,m\}$ implies $|V_0^*(\pi) - V_0^{\mathcal{D}}(\pi)| = |V_0^*(\pi) - V_0^*(d_i)| \leq \frac{1}{2}(\max_{\Gamma_0} - \min_{\Gamma_0})\,\mathrm{Diam}(\mathcal{D})$, since $V_0^{\mathcal{D}}(\pi) = V_0^{\mathcal{D}}(d_i) = V_0^*(d_i)$. This implies that $\boldsymbol{\rho}(V_0^*, V_0^{\mathcal{D}}) \leq \frac{1}{2}(\max_{\Gamma_0} - \min_{\Gamma_0})\,\mathrm{Diam}(\mathcal{D})$. For each $k \in \mathbf{N},\ i \in \{1,\ldots,m\},\ a \in \mathcal{A}$ note that $|\mathsf{h}(d_i, a, V_{k-1}^{\mathcal{D}}) - \mathsf{h}(d_i, a, V_{k-1}^*)| \leq \beta\boldsymbol{\rho}(V_{k-1}^*, V_{k-1}^{\mathcal{D}})$. This in turn implies $|V_k^{\mathcal{D}}(d_i) - V_k^*(d_i)| \leq \beta\boldsymbol{\rho}(V_{k-1}^*, V_{k-1}^{\mathcal{D}})$ from the corresponding definitions of $V_k^{\mathcal{D}}$ and V_k^*. Now V_k^* is Lipschitz continuous with a Lipschitz constant $\frac{C}{2}(\sum_{t=0}^{k-1}\beta^t) + \frac{1}{2}\beta^k(\max_{\Gamma_0} - \min_{\Gamma_0})$. By the definition of $V_k^{\mathcal{D}}$, for any $\pi \in \mathcal{D}_i$, $V_k^{\mathcal{D}}(\pi) = V_k^{\mathcal{D}}(d_i)$. Hence for any $\pi \in \mathcal{D}_i$,

$$
\begin{aligned}
&|V_k^{\mathcal{D}}(\pi) - V_k^*(\pi)| \\
&= |V_k^{\mathcal{D}}(d_i) - V_k^*(d_i) + V_k^*(d_i) - V_k^*(\pi)| \\
&\leq |V_k^{\mathcal{D}}(d_i) - V_k^*(d_i)| + |V_k^*(d_i) - V_k^*(\pi)| \\
&\leq \beta\boldsymbol{\rho}(V_{k-1}^{\mathcal{D}}, V_{k-1}^*) \quad + \\
&\qquad \left[\frac{C}{2}\left(\sum_{t=0}^{k-1}\beta^t\right) + \frac{1}{2}\beta^k(\max_{\Gamma_0} - \min_{\Gamma_0})\right] \mathrm{Diam}(\mathcal{D}).
\end{aligned}
$$

This implies

$$\boldsymbol{\rho}(V_k^{\mathcal{D}}, V_k^*)$$

$$\leq \quad \beta\boldsymbol{\rho}(V_{k-1}^{\mathcal{D}}, V_{k-1}^{*}) \quad + \\ \left[\frac{C}{2}\left(\sum_{t=0}^{k-1}\beta^t\right) + \frac{1}{2}\beta^k(\max_{\Gamma_0} - \min_{\Gamma_0})\right] \operatorname{Diam}(\mathcal{D}).$$

By an induction argument it easily follows that $\forall k \in \mathbf{N}$

$$\begin{aligned}
&\boldsymbol{\rho}(V_k^*, V_k^{\mathcal{D}}) \\
&\leq \quad \frac{C}{2}\left(\sum_{t=0}^{k-1}(t+1)\beta^t\right) \operatorname{Diam}(\mathcal{D}) \quad + \\
&\qquad \frac{(k+1)}{2}\beta^k(\max_{\Gamma_0} - \min_{\Gamma_0}) \operatorname{Diam}(\mathcal{D}) \\
&= \quad \frac{C}{2}\left(\sum_{t=0}^{k-1}\beta^t\left(\sum_{j=0}^{k-1-t}\beta^j\right)\right) \operatorname{Diam}(\mathcal{D}) \quad + \\
&\qquad \frac{(k+1)}{2}\beta^k(\max_{\Gamma_0} - \min_{\Gamma_0}) \operatorname{Diam}(\mathcal{D}).
\end{aligned}$$

When $\beta \in [0, 1)$,

$$\begin{aligned}
\sum_{t=0}^{k-1}\beta^t\left(\sum_{j=0}^{k-1-t}\beta^j\right) &\leq \sum_{t=0}^{k-1}\beta^t\frac{1}{1-\beta} \\
&\leq \frac{1}{(1-\beta)^2}.
\end{aligned}$$

Hence when $\beta \in [0, 1)$,

$$\begin{aligned}
\boldsymbol{\rho}(V_k^*, V_k^{\mathcal{D}}) \quad \leq \quad & \frac{C}{2}\frac{1}{(1-\beta)^2} \operatorname{Diam}(\mathcal{D}) \quad + \\
& \frac{(k+1)}{2}\beta^k(\max_{\Gamma_0} - \min_{\Gamma_0}) \operatorname{Diam}(\mathcal{D}).
\end{aligned}$$

When $\beta \in [0, 1)$, $V_k^{\mathcal{D}} \to V^{\mathcal{D}}$ as $k \to \infty$ uniformly on Δ since $\bar{V}_k^{\mathcal{D}} \to \bar{V}^{\mathcal{D}}$. Similarly, $V_k^* \to V^*$ as $k \to \infty$ uniformly on Δ. Hence taking the limit gives

$$\boldsymbol{\rho}(V^*, V^{\mathcal{D}}) \quad \leq \quad \frac{C}{2}\frac{1}{(1-\beta)^2} \operatorname{Diam}(\mathcal{D}),$$

since $(k+1)\beta^k \to 0$ as $k \to \infty$.

□

Proof of Theorem 2.3

For any $\pi \in \Delta$, $a \in \mathcal{A}$, and U, $V \in B(\Delta)$, $|\mathsf{h}(\pi, a, V) - \mathsf{h}(\pi, a, U)| \leq \beta\, \boldsymbol{\rho}(U, V)$. For $k \in \mathbf{N}$ this can be used to prove that $\boldsymbol{\rho}(H_{\delta_{k-1}} V^*_{k-1}, H\, V^*_{k-1}) \leq 2\beta\, \boldsymbol{\rho}(V^*_{k-1}, V^{\mathcal{D}}_{k-1})$. Also $\boldsymbol{\rho}(H_{\delta_{k-1}} V^{\Psi}_{k-1}, H_{\delta_{k-1}} V^*_{k-1}) \leq \beta \boldsymbol{\rho}(V^*_{k-1}, V^{\Psi}_{k-1})$. Hence

$$\begin{aligned}
&\boldsymbol{\rho}(V^*_k, V^{\Psi}_k) \\
&= \boldsymbol{\rho}(H\, V^*_{k-1}, H_{\delta_{k-1}} V^{\Psi}_{k-1}) \\
&\leq \boldsymbol{\rho}(H\, V^*_{k-1}, H_{\delta_{k-1}} V^*_{k-1}) + \boldsymbol{\rho}(H_{\delta_{k-1}} V^*_{k-1}, H_{\delta_{k-1}} V^{\Psi}_{k-1}) \\
&\leq 2\beta\, \boldsymbol{\rho}(V^*_{k-1}, V^{\mathcal{D}}_{k-1}) + \beta \boldsymbol{\rho}(V^*_{k-1}, V^{\Psi}_{k-1}).
\end{aligned}$$

Now $\boldsymbol{\rho}(V^*_0, V^{\Psi}_0) = 0$. Hence $\boldsymbol{\rho}(V^*_1, V^{\Psi}_1) \leq 2\beta\, \boldsymbol{\rho}(V^*_0, V^{\mathcal{D}}_0) \leq \beta\, (\max_{\Gamma_0} - \min_{\Gamma_0})\, \mathrm{Diam}(\mathcal{D})$. Using the bounds for $\boldsymbol{\rho}(V^*_k, V^{\mathcal{D}}_k)$ from Theorem 2.2, we may see by an induction argument that for $k > 1$,

$$\begin{aligned}
&\boldsymbol{\rho}(V^*_k, V^{\Psi}_k) \\
&\leq C \left(\sum_{t=1}^{k-1} \left(\sum_{j=1}^{t} j \right) \beta^t \right) \mathrm{Diam}(\mathcal{D}) \quad + \\
&\qquad \left(\sum_{t=1}^{k} t \right) \beta^k\, (\max_{\Gamma_0} - \min_{\Gamma_0})\, \mathrm{Diam}(\mathcal{D}) \\
&= C \left(\sum_{t=0}^{k-1} \frac{t(t+1)}{2} \beta^t \right) \mathrm{Diam}(\mathcal{D}) \quad + \\
&\qquad \frac{k(k+1)}{2} \beta^k\, (\max_{\Gamma_0} - \min_{\Gamma_0})\, \mathrm{Diam}(\mathcal{D}).
\end{aligned}$$

This proves Theorem 2.3. Observe that for $k > 1$, $\sum_{t=1}^{k-1} (\sum_{j=1}^{t} j) \beta^t = \sum_{t=1}^{k-1} t (\sum_{j=t}^{k-1} \beta^j)$. Hence when $\beta \in [0, 1)$,

$$\begin{aligned}
\sum_{t=1}^{k-1} \left(\sum_{j=1}^{t} j \right) \beta^t &\leq \frac{\beta}{1-\beta} \left(\sum_{t=0}^{k-2} (t+1)\, \beta^t \right) \\
&\leq \frac{\beta}{(1-\beta)} \frac{1}{(1-\beta)^2},
\end{aligned}$$

where the last inequality follows as in the proof of Theorem 2.2.

Hence $\lim_{k\to\infty} \boldsymbol{\rho}(V^*, V_k^{\Psi}) = \lim_{k\to\infty} \boldsymbol{\rho}(V_k^*, V_k^{\Psi}) \leq C\frac{\beta}{(1-\beta)^3}\mathrm{Diam}(\mathcal{D})$.

□

Proof of Theorem 2.4

For any $a \in \mathcal{A}$ and $k \in \mathbf{N}$, $\mathsf{h}(\pi, a, V_{k-1}^*)$ considered as a function of π is representable as the minimum of a finite number of linear functions on Δ and is Lipschitz continuous, and the same Lipschitz constant given in Theorem 2.1 for V_k^* holds. This fact may be obtained in a manner similar to that of the proof of Theorem 2.1. It may also be seen that for each $i \in \{1, \ldots, m\}$ and $k \in \mathbf{N}$, $|\mathsf{h}(d_i, \hat{\delta}_{k-1}(d_i), V_{k-1}^*) - V_k^*(d_i)| \leq 2\beta\, \boldsymbol{\rho}(V_{k-1}^*, V_{k-1}^{\mathcal{D}})$. Now for any $\pi \in \mathcal{D}_i$, $\hat{\delta}_k(\pi) = \hat{\delta}_k(d_i)$ by definition. Hence for any $\pi \in \mathcal{D}_i$,

$$
\begin{aligned}
&|V_k^{\hat{\Psi}}(\pi) - V_k^*(\pi)| \\
&= |\mathsf{h}(\pi, \hat{\delta}_{k-1}(\pi), V_{k-1}^{\hat{\Psi}}) - V_k^*(\pi)| \\
&\leq |\mathsf{h}(\pi, \hat{\delta}_{k-1}(\pi), V_{k-1}^{\hat{\Psi}}) - \mathsf{h}(\pi, \hat{\delta}_{k-1}(\pi), V_{k-1}^*)| \quad + \\
&\quad |\mathsf{h}(\pi, \hat{\delta}_{k-1}(\pi), V_{k-1}^*) - \mathsf{h}(d_i, \hat{\delta}_{k-1}(d_i), V_{k-1}^*)| \quad + \\
&\quad |\mathsf{h}(d_i, \hat{\delta}_{k-1}(d_i), V_{k-1}^*) - V_k^*(d_i)| \quad + \\
&\quad |V_k^*(d_i) - V_k^*(\pi)| \\
&\leq \beta\, \boldsymbol{\rho}(V_{k-1}^*, V_{k-1}^{\hat{\Psi}}) \quad + \\
&\quad 2\left(\frac{C}{2}(\sum_{t=0}^{k-1}\beta^t) + \frac{1}{2}\beta^k(\max_{\Gamma_0} - \min_{\Gamma_0})\right)\mathrm{Diam}(\mathcal{D}) \\
&\quad + 2\beta\, \boldsymbol{\rho}(V_{k-1}^*, V_{k-1}^{\mathcal{D}}).
\end{aligned}
$$

Hence

$$\begin{aligned}
&\boldsymbol{\rho}(V_k^{\hat{\Psi}}, V_k^*) \\
&\leq \beta\, \boldsymbol{\rho}(V_{k-1}^*, V_{k-1}^{\hat{\Psi}}) \quad + \\
&\quad 2\left(\frac{C}{2}(\sum_{t=0}^{k-1}\beta^t) + \frac{1}{2}\beta^k(\max_{\Gamma_0} - \min_{\Gamma_0})\right)\mathrm{Diam}(\mathcal{D}) \\
&\quad + 2\beta\, \boldsymbol{\rho}(V_{k-1}^*, V_{k-1}^{\mathcal{D}}).
\end{aligned}$$

Now $\boldsymbol{\rho}(V_0^*, V_0^{\hat{\Psi}}) = 0$. Using the bounds for $\boldsymbol{\rho}(V_k^*, V_k^{\mathcal{D}})$ from Theorem 2.2, we may see by an induction argument that for $k \in \mathbf{N}$,

$$\begin{aligned}
&\boldsymbol{\rho}(V_k^*, V_k^{\hat{\Psi}}) \\
&\leq C\left(\sum_{t=0}^{k-1}\left(\sum_{j=0}^{t}(j+1)\right)\beta^t\right)\mathrm{Diam}(\mathcal{D}) \quad + \\
&\quad \left(\sum_{t=1}^{k}(t+1)\right)\beta^k\,(\max_{\Gamma_0} - \min_{\Gamma_0})\,\mathrm{Diam}(\mathcal{D}) \\
&= C\left(\sum_{t=0}^{k-1}\frac{(t+1)(t+2)}{2}\beta^t\right)\mathrm{Diam}(\mathcal{D}) \quad + \\
&\quad \frac{k(k+3)}{2}\beta^k\,(\max_{\Gamma_0} - \min_{\Gamma_0})\,\mathrm{Diam}(\mathcal{D}).
\end{aligned}$$

When $\beta \in [0, 1)$, it may be seen in a manner similar to that in the proof of Theorem 2.3, that for $k \in \mathbf{N}$,

$$\sum_{t=0}^{k-1}\left(\sum_{j=0}^{t}(j+1)\right)\beta^t \leq \frac{1}{(1-\beta)^3}.$$

Hence $\lim_{k\to\infty}\boldsymbol{\rho}(V^*, V_k^{\hat{\Psi}}) = \lim_{k\to\infty}\boldsymbol{\rho}(V_k^*, V_k^{\hat{\Psi}}) \leq C\frac{1}{(1-\beta)^3}\mathrm{Diam}(\mathcal{D})$.

□

2.6 An Example For a Non-Lipschitz Bayesian Transition Function

We give an example to show that the Bayesian transition function $F(\pi, a, o)$ is not necessarily a Lipschitz continuous function of π for fixed action a and a fixed observation o. This implies that Assumption A.2 of [9] need not be satisfied in general for a POMDP, and hence the results given in [9] cannot be adapted directly to our case. Consider a POMDP with $\mathcal{S} = \{1, 2, 3\}$; $\mathcal{O} = \{1, 2\}$; $\mathcal{A} = \{1\}$. Let

$$P(1) = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 1 \end{bmatrix},$$

$q(i, 1, 1, 1) = q(i, 1, 1, 2) = \frac{1}{2}$; $q(i, 1, 2, 1) = q(i, 1, 2, 2) = \frac{1}{2}$; $q(i, 1, 3, 1) = 0$, $q(i, 1, 3, 2) = 1$; for all $i \in \{1, \ldots, n\}$. In this case it may be seen that

$$[F(\pi, 1, 1)]_2 = \frac{\frac{1}{4}\pi_1}{\frac{2}{4}\pi_1 + \frac{1}{4}\pi_2}$$

and

$$\frac{\partial [F(\pi, 1, 1)]_2}{\partial \pi_2} = \frac{-\frac{1}{16}\pi_1}{\left(\frac{2}{4}\pi_1 + \frac{1}{4}\pi_2\right)^2}.$$

Let $\pi = (\alpha, \alpha, 1 - 2\alpha)$ with $\alpha \in (0, \frac{1}{2})$. Then

$$\left.\frac{\partial [F(\pi, 1, 1)]_2}{\partial \pi_2}\right|_{\pi=(\alpha,\alpha,1-2\alpha)} = -\frac{1}{9}\frac{1}{\alpha}.$$

Now as $\alpha \to 0^+$ this quantity tends to $-\infty$. This implies that $F(\pi, 1, 1)$ is not Lipschitz continuous.

Chapter 3

A Stochastic Approximation Algorithm For Periodic Markov Processes

In this chapter we discuss a stochastic approximation algorithm which is a slight generalization of the results in [8]. We allow the Markov process to be periodic (i.e. it need not be aperiodic). The discussion of this chapter closely follows that of Chapter 1 in part II of [8]. We use the same notations as in Chapter 1, Part II of [8]. The notations in this chapter are self contained. Consider the algorithm

$$\theta_{n+1} \quad = \quad \theta_n + \gamma_{n+1} H(\theta_n, X_{n+1}) + \gamma_{n+1}^2 \rho_{n+1}(\theta_n, X_{n+1}) \tag{3.1}$$

where θ_n evolves in $\mathbf{R}^d$ and the state vector X_n lies in $\mathbf{R}^k$ or in a subset of $\mathbf{R}^k$, say $\mathcal{X}$. H and ρ_n are two functions from $\mathbf{R}^d \times \mathbf{R}^k$ to $\mathbf{R}^d$. We assume that the random variables (r.v.) $\theta_0, X_0, X_1, \ldots, X_n, \ldots$ are defined on a probability space $(\Omega, \mathcal{F}, P)$, and we denote the σ-field of events generated by the r.v. $\theta_0, X_0, \ldots, X_n$ by $\mathcal{F}_n$. Let $\mathbf{N}$ denote the set of natural numbers, i.e. the set of positive integers. In all that follows, the following assumptions are made:

(A.1) $(\gamma_n)_{n \in \mathbf{N}}$ is a sequence of non-negative real numbers such that $\sum_n \gamma_n = +\infty$ and $\sum_{n=1}^{\infty} \mid \gamma_{n+1} - \gamma_n \mid \equiv \tilde{K}_\gamma < \infty$. Then $\tilde{\gamma} \equiv \sup_{n \in \mathbf{N}} \gamma_n < \infty$.

□

Denote $K_\gamma^n = \sum_{k=n}^{\infty} \mid \gamma_k - \gamma_{k+1} \mid$. Then $K_\gamma^1 = \tilde{K}_\gamma$. Note that $\lim_{n \to \infty} K_\gamma^n = 0$.

(A.2) There exists a family $\{\Pi_\theta : \theta \in \mathbf{R}^d\}$ of transition probabilities $\Pi_\theta(x, A)$ on $\mathbf{R}^k$ such that, for any Borel subset A of $\mathbf{R}^k$, we have

$$P[X_{n+1} \in A \mid \mathcal{F}_n] \;=\; \Pi_{\theta_n}(X_n, A) \tag{3.2}$$

□

Assumption (A.2) says that the 2-tuple $(X_n, \theta_n)_{n\geq 0}$ is a Markov process. Its transition probability depends on n (since γ_n and ρ_n depend on n). It is therefore an inhomogeneous Markov process. Note that Assumption (A.1) is different than in [8] and does not need (γ_n) to be a non-increasing sequence.

Notation.

a. Let $P_{x,a}$ denote the distribution of $(X_n, \theta_n)_{n\geq 0}$ for the initial conditions $X_0 = x, \theta_0 = a$.

b. If, more precisely,

$$P_{x,a}^{(\gamma_n, \rho_n; n\geq 0)}$$

denotes the distribution of $(X_n, \theta_n)_{n\geq 0}$ for the given sequence $(\gamma_n, \rho_n)_{n\geq 0}$, with initial conditions $X_0 = x, \theta_0 = a$, then the conditional distribution of $(X_{n+k}, \theta_{n+k})_{k\geq 0}$ given $\mathcal{F}_n$ is

$$P_{X_n, \theta_n}^{(\gamma_{n+k}, \rho_{n+k}; k\geq 0)}$$

c. In what follows, it will be useful to express the trajectory of the algorithm $n \to \theta_n$ in the form of a continuous-time process. To this end, we set

$$t_0 \;=\; 0,\; t_1 = \gamma_1,\; \ldots,\; t_n = \sum_{i=1}^{n} \gamma_i \tag{3.3}$$

$$\theta(t) \quad = \quad \sum_{k\geq 0} I(t_k \leq t < t_{k+1})\theta_k \tag{3.4}$$

where $I(A)$ denotes the characteristic function of the set A (often denoted by 1_A).

The study of the behaviour of $\theta(t)$ between times t_n and $t_n + T$ thus reduces to the study of the behaviour of θ_k for integers k between n and $m(n,T)$, where

$$m(n,T) \quad = \quad \inf\{k : k \geq n, \gamma_{n+1} + \ldots + \gamma_{k+1} \geq T\} \tag{3.5}$$

For simplicity, we shall denote

$$m(T) \quad = \quad m(0,T) \tag{3.6}$$

d. For any function $f(x,\theta)$ on $\mathbf{R}^k \times \mathbf{R}^d$, we shall denote the partial mapping $x \to f(x,\theta)$ by f_θ. In particular, $\Pi_\theta f_\theta$ denotes the function

$$x \quad \to \quad \int f(y,\theta)\Pi_\theta(x,dy)$$

Similarly $\Pi_\theta^k f_\theta$ for $k \geq 1$ denotes the function

$$x \quad \to \quad \int (\Pi_\theta^{k-1} f_\theta(y))\Pi_\theta(x,dy)$$

with $\Pi_\theta^0 f_\theta \equiv f_\theta$.

e. For a real number α, $|\alpha|$ denotes the absolute value of α. For a vector v, $|v|$ denotes the ℓ_2 norm (Eucledian norm) of v. For a matrix A, $|A|$ denotes the matrix norm induced by the ℓ_2 norm [25].

3.1 General Assumptions On H, ρ_n And Π

We shall frequently denote the function $x \to H(\theta, x)$ by H_θ. We shall assume that D is an open subset of $\mathbf{R}^d$. The functions H and ρ_n will be required to satisfy:

(A.3) For any compact subset Q of D, there exist constants C_1, C_2, q_1, q_2 (depending on Q), such that for all $\theta \in Q$, and all n we have

(i)

$$|H(\theta, x)| \leq C_1(1 + |x|^{q_1})$$

(ii)

$$|\rho_n(\theta, x)| \leq C_2(1 + |x|^{q_2})$$

□

When we wish to express the dependence on Q explicitly in the above formulae, we shall write $C_i(Q)$ or $q_i(Q)$.

The verification of the fundamental assumption which we shall introduce next is central to the study of the algorithm. Note that this assumption is slightly different from that of [8, page 216] to take into account the periodicity of the Markov process.

(A.4) There exists a positive integer p such that the state space $\mathcal{X}$ can be partitioned into disjoint Borel sets $\mathcal{X}_0, \cdots, \mathcal{X}_{p-1}$ with $\Pi_\theta(x, \mathcal{X}_{(i+1) \bmod p}) = 1 \ \forall x \in \mathcal{X}_i, \theta \in D$. Further there exist functions $h_0, \cdots, h_{p-1}$ on D, and for each $\theta \in D$ a function $\nu_\theta(\cdot)$ on $\mathcal{X}$ such that

(i) h_i is locally Lipschitz on D for $i = 0, \cdots, p-1$

(ii) $(I - \Pi_\theta)\nu_\theta = \tilde{H}_\theta$ for all $\theta \in D$, where $\tilde{H}_\theta(x) = H_\theta(x) - h_{\tilde{p}(x)}(\theta)$ with $\tilde{p}(x) = i$, if $x \in \mathcal{X}_i$.

(iii) for all compact subsets Q of D, there exist constants $C_3, C_4, q_3, q_4, \lambda \in [\frac{1}{2}, 1]$, such that for all $\theta, \theta' \in Q$

$$|\nu_\theta(x)| \leq C_3(1 + |x|^{q_3}) \tag{3.7}$$

$$|\Pi_\theta \nu_\theta(x) - \Pi_{\theta'}\nu_{\theta'}(x)| \leq C_4|\theta - \theta'|^\lambda(1 + |x|^{q_4}) \tag{3.8}$$

□

Let $h(\theta) \equiv \frac{1}{p}(h_0(\theta) + \cdots + h_{p-1}(\theta))$. Let

$$\tilde{L}(Q) \equiv \max\{\tilde{L}_0(Q), \cdots, \tilde{L}_{p-1}(Q)\}$$

where $\tilde{L}_i(Q)$ is the Lipschitz constant for $h_i(\cdot)$ on Q. Let

$$\tilde{M}(Q) = \max_{i \in \{0,\cdots,p-1\}} \sup_{\theta \in Q} |h_i(\theta)|$$

Comments on (A.4)

Note that the functions $H_\theta,\ h(\theta),\ \nu(\theta),\ \tilde{H}_\theta$ and $h_i(\theta)$ take their values in $\mathbf{R}^d$. Condition (A.4-ii) implies that for each $i = 1, \ldots, d$

$$(I - \Pi_\theta)\nu_\theta^i(x) = \tilde{H}_\theta^i(x), \quad \forall x \in \mathcal{X}$$

where the superscript i denotes the i^{th} coordinate in $\mathbf{R}^d$.

Concerning the importance of (A.4), note that if for all θ, the Markov process with transition probability Π_θ is positive recurrent, with invariant distibution Γ_θ,

and if we set

$$h_i(\theta) \;=\; p \int_{\mathcal{X}_i} H_\theta(y)\Gamma_\theta(dy) \qquad i = 0, \ldots, p-1 \tag{3.9}$$

then

$$h(\theta) \;=\; \int_{\mathcal{X}} H_\theta(y)\Gamma_\theta(dy) \tag{3.10}$$

(or more concisely $\Gamma_\theta H_\theta$), and the function $\tilde{H}_\theta$ has the property that for each $i = 0, \ldots, p-1$,

$$\int_{\mathcal{X}_i} \tilde{H}_\theta(y)\Gamma_\theta(dy) = 0$$

since

$$\int_{\mathcal{X}_i} \Gamma_\theta(dy) = \frac{1}{p} \qquad i = 0, \ldots, p-1$$

and thus equation (A.4-ii) has a solution ν_θ. Moreover in most cases, this solution may be expressed in the form

$$\nu_\theta(y) \;=\; \sum_{k \geq 0} \Pi_\theta^k \tilde{H}_\theta(y) \tag{3.11}$$

when the series is convergent.

The Local Boundedness assumption

(A.5) For any compact subset Q of D and any $q > 0$, there exists $\mu_q(Q) < \infty$ such that for all $n, x \in \mathbf{R}^k, a \in \mathbf{R}^d$

$$E_{x,a}\{I(\theta_k \in Q, k \leq n)(1 + |X_{n+1}|^q)\} \;\leq\; \mu_q(Q)(1 + |x|^q) \tag{3.12}$$

□

Remarks on Assumption (A.5)

1. If the inequality (3.12) is true for q, then it is true for $q' < q$ [8, page 220].

2. In the definition of (A.5) the inequality is assumed for all $q > 0$. In fact, for the proofs we need only a weaker assumption, namely that the inequality (3.12) is valid for a sufficiently large q, i.e. larger than a well-defined function of the exponents q_i in (A.3) and (A.4) (refer Proposition 3.1).

3. Without loss of generality we assume that μ_q is an increasing (non-decreasing) function of q.

3.2 Decomposition Of The General Algorithm

3.2.1

When $\tilde{\gamma}$ tends to zero, the algorithm $\theta(t)$ has a tendency to follow the solution of the differential equation (deterministic) with initial condition $a = \bar{\theta}(0)$.

$$\bar{\theta}'(t) = h(\bar{\theta}(t)) \tag{3.13}$$

This is because $\bar{\theta}(t_n)$ is close (Euler's approximation) to the solution $\bar{\theta}_n$ of

$$\begin{aligned} \bar{\theta}_{n+1} &= \bar{\theta}_n + \gamma_{n+1} h(\bar{\theta}_n) \\ \bar{\theta}_0 &= a \end{aligned} \tag{3.14}$$

and because Algorithm 3.1 (or equation 3.1) may be written in the form

$$\theta_{n+1} = \theta_n + \gamma_{n+1} h(\theta_n) + \varepsilon_n \tag{3.15}$$

where

$$\begin{aligned} \varepsilon_n &= \theta_{n+1} - \theta_n - \gamma_{n+1} h(\theta_n) \\ &= \gamma_{n+1}[H(\theta_n, X_{n+1}) - h(\theta_n) + \gamma_{n+1}\rho_{n+1}(\theta_n, X_{n+1})] \end{aligned} \tag{3.16}$$

is a small fluctuation for small $\tilde{\gamma}$. We desire to obtain upper bounds on the fluctuations ε_n. More generally, in the sequel we shall require upper bounds for the expressions

$$\varepsilon_n(\phi) = \phi(\theta_{n+1}) - \phi(\theta_n) - \gamma_{n+1}\phi'(\theta_n) \cdot h(\theta_n) \tag{3.17}$$

Let ϕ be a C^2 function (i.e. having continuous second partial derivatives) from $\mathbf{R}^d$ to $\mathbf{R}$ with bounded second derivatives. For the compact subset Q of D we denote

$$\left.\begin{aligned} M_0(Q) &= \sup_{\theta \in Q} |\phi(\theta)| \\ M_1(Q) &= \sup_{\theta \in Q} |\phi'(\theta)| \\ M_2(Q) &= \sup_{\theta \in Q} |\phi''(\theta)| \\ M_2 &= \sup_{\theta \in \mathbf{R}^d} |\phi''(\theta)| \end{aligned}\right\} \tag{3.18}$$

Here ϕ' is the gradient of ϕ and ϕ'' is the Hessian, $\left[\frac{\partial^2 \phi(\theta)}{\partial \theta^i \partial \theta^j}\right]$ of ϕ at θ [25]. Then there exists a matrix $\tilde{R}(\phi, \theta, \theta')$ by Taylor's formula [3], such that

$$\phi(\theta') - \phi(\theta) - (\theta' - \theta) \cdot \phi'(\theta) = \underbrace{(\theta - \theta')\tilde{R}(\phi, \theta, \theta')(\theta - \theta')}_{R(\phi, \theta, \theta')} \tag{3.19}$$

with, for all $\theta, \theta' \in \mathbf{R}^d$

$$|R(\phi, \theta, \theta')| \leq M_2 |\theta' - \theta|^2 \tag{3.20}$$

Thus for all k

$$\begin{aligned}\varepsilon_k(\phi) &= \phi'(\theta_k)\cdot[(\theta_{k+1}-\theta_k)-\gamma_{k+1}h(\theta_k)]+R(\phi,\theta_k,\theta_{k+1})\\ &= \gamma_{k+1}\phi'(\theta_k)\cdot[H(\theta_k,X_{k+1})-h(\theta_k)]\\ &\quad+\underbrace{\left(\gamma_{k+1}^2\phi'(\theta_k)\cdot\rho_{k+1}(\theta_k,X_{k+1})+R(\phi,\theta_k,\theta_{k+1})\right)}_{A_k^1}\end{aligned}\tag{3.21}$$

with

$$\begin{aligned}&|R(\phi,\theta_k,\theta_{k+1})|\\ &\leq \gamma_{k+1}^2M_2|H(\theta_k,X_{k+1})+\gamma_{k+1}\rho_{k+1}(\theta_k,X_{k+1})|^2\end{aligned}\tag{3.22}$$

3.2.2 Decomposition Of $\varepsilon_n(\phi)$

Using (A.4-ii) we may write (3.21) as

$$\begin{aligned}&\varepsilon_k(\phi)\\ &= \phi(\theta_{k+1})-\phi(\theta_k)-\gamma_{k+1}\phi'(\theta_k)\cdot h(\theta_k)\\ &= \gamma_{k+1}\phi'(\theta_k)[H(\theta_k,X_{k+1})-h_{\tilde{p}(X_{k+1})}(\theta_k)-h(\theta_k)+h_{\tilde{p}(X_{k+1})}(\theta_k)]+A_k^1\\ &= \gamma_{k+1}\phi'(\theta_k)[\nu_{\theta_k}(X_{k+1})-\Pi_{\theta_k}\nu_{\theta_k}(X_{k+1})-h(\theta_k)+h_{\tilde{p}(X_{k+1})}(\theta_k)]+A_k^1\\ &= \gamma_{k+1}\phi'(\theta_k)\cdot[\nu_{\theta_k}(X_{k+1})-\Pi_{\theta_k}\nu_{\theta_k}(X_k)]\\ &\quad+\gamma_{k+1}\phi'(\theta_k)[\Pi_{\theta_k}\nu_{\theta_k}(X_k)-\Pi_{\theta_k}\nu_{\theta_k}(X_{k+1})]\\ &\quad+\gamma_{k+1}\phi'(\theta_k)[h_{\tilde{p}(X_{k+1})}(\theta_k)-h(\theta_k)]+A_k^1\\ &= A_k^2+A_k^3+A_k^4+A_k^1\end{aligned}$$

This calculation makes sense only when $\theta_k\in D$, since h is only defined on D.

Hence we introduce for a fixed compact subset Q of D

$$\tau \;=\; \tau(Q) \;=\; \inf(n : \theta_n \notin Q) \tag{3.23}$$

Let

$$\psi_\theta(x) \;=\; \phi'(\theta) \cdot \Pi_\theta \nu_\theta(x) \tag{3.24}$$

Then in $\{\tau \geq n\}$ and for $r < n$ we have :

$$\begin{aligned}
\sum_{k=r}^{n-1} \varepsilon_k(\phi) &= \sum_{k=r}^{n-1}(A_k^1 + A_k^2 + A_k^4) + \sum_{k=r}^{n-1} \gamma_{k+1}(\psi_{\theta_k}(X_k) - \psi_{\theta_k}(X_{k+1})) \\
&= \sum_{k=r}^{n-1}(A_k^1 + A_k^2 + A_k^4) \\
&\quad + \sum_{k=r+1}^{n-1} \gamma_{k+1}(\psi_{\theta_k}(X_k) - \psi_{\theta_{k-1}}(X_k)) \\
&\quad + \sum_{k=r+1}^{n-1} (\gamma_{k+1} - \gamma_k)\psi_{\theta_{k-1}}(X_k) \\
&\quad + \gamma_{r+1}\psi_{\theta_r}(X_r) - \gamma_n \psi_{\theta_{n-1}}(X_n)
\end{aligned}$$

We have the following lemma

Lemma 3.1 *For $r < n$ in $\{n \leq \tau\}$ we have*

$$\begin{aligned}
\sum_{k=r}^{n-1} \varepsilon_k(\phi) &= \sum_{k=r}^{n-1} \varepsilon_k^{(1)} + \sum_{k=r+1}^{n-1} \varepsilon_k^{(2)} + \sum_{k=r+1}^{n-1} \varepsilon_k^{(3)} + \sum_{k=r}^{n-1} \varepsilon_k^{(4)} \\
&\quad + \sum_{k=r}^{n-1} \varepsilon_k^{(5)} + \eta_{n;r}
\end{aligned}$$

where

$$\begin{aligned}
\varepsilon_k^{(1)} &= \gamma_{k+1}\phi'(\theta_k) \cdot (\nu_{\theta_k}(X_{k+1}) - \Pi_{\theta_k}\nu_{\theta_k}(X_k)) \\
\varepsilon_k^{(2)} &= \gamma_{k+1}(\psi_{\theta_k}(X_k) - \psi_{\theta_{k-1}}(X_k))
\end{aligned}$$

$$
\begin{aligned}
\varepsilon_k^{(3)} &= (\gamma_{k+1}-\gamma_k)\psi_{\theta_{k-1}}(X_k) \\
\varepsilon_k^{(4)} &= \gamma_{k+1}^2\phi'(\theta_k)\cdot\rho_{k+1}(\theta_k,X_{k+1})+R(\phi,\theta_k,\theta_{k+1}) \\
\varepsilon_k^{(5)} &= \gamma_{k+1}\phi'(\theta_k)[h_{\tilde{p}(X_{k+1})}(\theta_k)-h(\theta_k)] \\
\eta_{n;r} &= \gamma_{r+1}\psi_{\theta_r}(X_r)-\gamma_n\psi_{\theta_{n-1}}(X_n)
\end{aligned}
$$

□

Remark 3.1 *Using (A.5) and (A.4-iii) we get*

$$
\begin{aligned}
|\Pi_\theta\nu_\theta(x)| &= |E_{x,\theta}(\nu_\theta(X_1))| \\
&\le E_{x,\theta}|\nu_\theta(X_1)| \\
&\le C_3E_{x,\theta}(1+|X_1|^{q_3}) \\
&\le C_3\mu_{q_3}(1+|x|^{q_3})
\end{aligned}
$$

for all $\theta\in Q$*, i.e.*

$$
\sup_{\theta\in Q}|\Pi_\theta\nu_\theta(x)| \le C_3\mu_{q_3}(1+|x|^{q_3})
$$

□

Remark 3.2 *From (3.7), (3.8) and (3.18) we have*

$$
\sup_{\theta\in Q}|\psi_\theta(x)| \le M_1C_3\mu_{q_3}(1+|x|^{q_3}) \tag{3.25}
$$

$$
\begin{aligned}
\sup_{\theta,\theta'\in Q}|\psi_\theta(x)-\psi_{\theta'}(x)| &\le M_1C_4(1+|x|^{q_4})|\theta-\theta'|^\lambda \\
&\quad +M_2C_3\mu_{q_3}(1+|x|^{q_3})|\theta-\theta'|
\end{aligned} \tag{3.26}
$$

□

3.3 L^2 Estimates

The aim of this section is to prove Proposition 3.1 (below), which gives a mean squares upper bound for the "fluctuation"

$$\sup_{n\leq m\wedge\tau}\left|\sum_{k=0}^{n-1}\varepsilon_k(\phi)\right|$$

where τ is the time at which the process θ_n leaves the compact subset Q.

In this section, Q is a fixed compact set. The "constants" which appear in the results may depend upon Q just as they depend upon the parameters C_i, μ_q and λ of the assumptions and upon the numbers $M_i(\phi)$ associated with the given function ϕ (cf. (3.18)). On the other hand they are valid for all non-negative sequences $(\gamma_n)_{n\geq 1}$ such that $\sum_{k=1}^{\infty}|\gamma_k-\gamma_{k+1}|\leq\tilde{K}_\gamma<+\infty$. Let $\tilde{\gamma}\equiv\sup_k\gamma_k<+\infty$.

3.3.1

We state the following lemmas from [8, pages 224-228]

Lemma 3.2 *There exists a constant A_1 such that:*

$$E_{x,a}\{\sup_{n\leq m}I(n\leq\tau)\left|\sum_{k=0}^{n-1}\varepsilon_k^{(1)}\right|^2\}\ \leq\ A_1(1+|x|^{2q_3})\sum_{k=0}^{m-1}\gamma_{k+1}^2$$

where using the constants of Assumptions (A.3) and (A.4)

$$A_1\leq\tilde{A}_1\mu_{2q_3}(Q)M_1^2(Q)C_3^2(Q)$$

the constant $\tilde{A}_1$ being independent of Q. Moreover on $\{\tau=+\infty\}$, $\sum_{k=0}^{n-1}\varepsilon_k^{(1)}$ converges a.s. and in L^2 if $\sum_{k=0}^{\infty}\gamma_{k+1}^2<\infty$. □

In considering the following terms we note that for $i = 2, 3, 4$

$$\begin{aligned} E\{\sup_{n\le m} I(n \le \tau)\left|\sum_{k=0}^{n-1} \varepsilon_k^{(i)}\right|^2\} &\le E\left(\sum_{k=0}^{m\wedge\tau-1} |\varepsilon_k^{(i)}\|\right)^2 \\ &= E\left(\sum_{k=0}^{m-1} |\varepsilon_k^{(i)}| I(k+1 \le \tau)\right)^2 \end{aligned}$$

with the convention: $\varepsilon_0^{(2)} = \varepsilon_0^{(3)} = 0$.

Lemma 3.3 *If $\tilde{\gamma} \le 1$, then there exists a constant A_2 such that for all m:*

$$E_{x,a}\left\{\sum_{k=1}^{m\wedge\tau-1} |\varepsilon_k^{(2)}|\right\}^2 \le A_2(1+|x|^{s_1})(\sum_{k=0}^{m-1} \gamma_{k+1}^{1+\lambda})^2$$

with $s_1 = \max(2q_4 + 2\lambda(q_1 \vee q_2), 2q_3 + 2(q_1 \vee q_2))$, and using the constants of (A.3), (A.4) and (A.5) and denoting $C_1(Q) + \tilde{\gamma}C_2(Q)$ by $\bar{C}(Q)$:

$$A_2 \le \tilde{A}_2\, \mu_{s_1}(Q)\, \max\{1, \mu_{q_3}^2(Q)\}\, [\bar{C}^{2\lambda}(Q) M_1^2(Q) C_4^2(Q) + \bar{C}^2(Q) M_2^2(Q) C_3^2(Q)]$$

$\tilde{A}_2$ being a constant independent of Q.

□

Note that [8, Lemma 3, page 225] γ_1 is replaced by $\tilde{\gamma}$ in the definition of $\bar{C}(Q)$ in the statement of Lemma 3.3. Also we have an additional term $\max\{1, \mu_{q_3}^2(Q)\}$ in the bound for A_2 (which was inadvertently omitted in [8]). The restriction that $\tilde{\gamma} \le 1$ can be removed if we allow $\tilde{A}_2$ to be dependent on $\tilde{\gamma}$.

Lemma 3.4 *There exists a constant A_3, such that for all n*

$$E_{x,a}\left\{\sum_{k=1}^{m\wedge\tau-1} |\varepsilon_k^{(3)}|\right\}^2 \le A_3(1+|x|^{2q_3})\tilde{K}_\gamma^2$$

with

$$A_3 \le \tilde{A}_3 M_1^2(Q) C_3^2(Q) \mu_{2q_3}^3(Q)$$

$\tilde{A}_3$ *being a constant independent of* Q.

□

Note that Lemma 3.4 is slightly different from Lemma 4 of [8, page 226] in that γ_1 is replaced by $\tilde{K}_\gamma$.

Lemma 3.5 *Denote* $s_2 = \sup(4q_1, 4q_2)$. *There exists a constant* A_4 *such that for all* m

$$E_{x,a}\left\{\sum_{k=0}^{m\wedge\tau-1}|\varepsilon_k^{(4)}|\right\}^2 \;\leq\; A_4(1+|x|^{s_2})(\sum_{k=0}^{m-1}\gamma_{k+1}^2)^2$$

with

$$A_4 \;\leq\; \tilde{A}_4\mu_{s_2}(Q)[C_2^2(Q)M_1^2(Q) \;+\; C_1^4(Q) \;+\; \tilde{\gamma}^4C_2^4(Q)]$$

$\tilde{A}_4$ *being a constant independent of* Q.

□

Note that [8, Lemma 5, page 227] γ_1 is replaced by $\tilde{\gamma}$ in the statement of Lemma 3.5.

Lemma 3.6 *There exists a constant* A_5 *such that*

$$E_{x,a}\{\sup_{1\leq n\leq m} I(n\leq\tau)\,|\eta_{n;0}|^2\} \;\leq\; A_5(1+|x|^{2q_3})\sum_{k=0}^{m-1}\gamma_{k+1}^2$$

with

$$A_5 \;\leq\; \tilde{A}_5M_1^2(Q)C_3^2(Q)\mu_{2q_3}^3(Q)$$

$\tilde{A}_5$ *being independent of* Q. *Moreover* $\eta_{n;0}$ *converges a.s. and in* L^2 *on* $\{\tau=+\infty\}$ *when* $\sum_{k=0}^{\infty}\gamma_{k+1}^2<\infty$. □

Note that [8, Lemma 6, page 227] $\mu_{2q_3}^2$ is replaced by $\mu_{2q_3}^3$ in the statement of Lemma 3.6.

The following lemma is new.

Lemma 3.7 *There exist constants A_6 and A_7 such that*

$$E_{x,a}\{\sup_{n\le m} I(n\le\tau)\left|\sum_{k=0}^{n-1}\varepsilon_k^{(5)}\right|^2\} \;\le\; A_6[\tilde{K}_\gamma^2+\sum_{k=0}^{m-1}\gamma_{k+1}^2]+A_7(1+|x|^{s_3})(\sum_{k=0}^{m-1}\gamma_{k+1}^2)^2$$

with $s_3=2(q_1\vee q_2)$, $A_6\le\tilde{A}_6M_1^2(Q)\tilde{M}^2(Q)$ and $A_7\le\tilde{A}_7\mu_{s_3}(Q)\bar{C}^2(Q)[M_1^2(Q)\tilde{L}^2(Q)+M_2^2\tilde{M}^2(Q)]$. Here $\bar{C}(Q)=[C_1(Q)+\tilde{\gamma}C_2(Q)]$ and $\tilde{A}_6$ and $\tilde{A}_7$ are constants independent of Q. Moreover, on $\{\tau=+\infty\}$, $\sum_{k=0}^{n-1}\varepsilon_k^{(5)}$ converges a.s. and in L^2 if $\sum\gamma_{k+1}^2<\infty$.

□

The proofs of Lemmas 3.2, 3.3, 3.5 and Lemma 3.6 are given in [8]. The proof of Lemma 3.2 uses L^2 maximal inequality and the L^2 convergence theorem of martingales [21, pages 248-249].

The proof of Lemma 3.4 is almost similar to that in [8], but is given below.

Proof of Lemma 3.4

Using (3.25) we obtain

$$\begin{aligned}
&E_{x,a}\left\{\sum_{k=1}^{m\wedge\tau-1}|\varepsilon_k^{(3)}|\right\}^2\\
&\le\; KE_{x,a}\left\{\sum_{k=1}^{m-1}|(\gamma_k-\gamma_{k+1})|(1+|X_k|^{q_3})I(k+1\le\tau)\right\}^2\\
&\le\; K\sum_{k=1}^{m-1}|(\gamma_k-\gamma_{k+1})|\sum_{k=1}^{m-1}|(\gamma_k-\gamma_{k+1})|E_{x,a}\{(1+|X_k|^{q_3})^2I(k+1\le\tau)\}
\end{aligned}$$

with $K \leq M_1^2(Q)C_3^2(Q)\mu_{q_3}^2(Q)$. Thus from (A.5)

$$E_{x,a}\left\{\sum_{k=1}^{m\wedge\tau-1}|\varepsilon_k^{(3)}|\right\}^2 \leq A_3(1+|x|^{2q_3})\tilde{K}_\gamma^2$$

with $A_3 \leq \tilde{A}_3 M_1^2(Q)C_3^2(Q)\mu_{2q_3}^3(Q)$.

□

Next we embark on proving Lemma 3.7.

Proof of Lemma 3.7

Let $0 \leq n \leq m \wedge \tau - 1$.

$$\begin{aligned}\left|\sum_{k=0}^{n}\varepsilon_k^{(5)}\right| &\leq I_{\{\lfloor\frac{n+1}{p}\rfloor\geq 1\}}\sum_{k=0}^{\lfloor\frac{n+1}{p}\rfloor-1}\left|\sum_{l=0}^{p-1}\varepsilon_{(kp+l)}^{(5)}\right| \\ &\quad + I_{\{\lfloor\frac{n+1}{p}\rfloor p\neq(n+1)\}}\underbrace{\sum_{k=\lfloor\frac{n+1}{p}\rfloor p}^{n}|\varepsilon_k^{(5)}|}_{\text{atmost } p-1 \text{ terms}}\end{aligned}$$

Here $\lfloor\alpha\rfloor$ denotes the floor of the real number α. Note that $\varepsilon_k^{(5)} = 0$ when $p = 1$.

Now let j be an integer such that $\theta_j, \theta_{j+1}, \cdots, \theta_{j+p-1}$ are in the set Q. Then

$$\begin{aligned}\sum_{l=0}^{p-1}\varepsilon_{j+l}^{(5)} &= \sum_{l=0}^{p-1}\gamma_{j+l+1}\phi'(\theta_{j+l})\left[h_{\tilde{p}(X_{j+l+1})}(\theta_{j+l}) - h(\theta_{j+l})\right] \\ &= \sum_{l=0}^{p-1}(\gamma_{j+1}\phi'(\theta_j) + (\gamma_{j+l+1}-\gamma_{j+1})\phi'(\theta_j) \\ &\quad +\gamma_{j+l+1}(\phi'(\theta_{j+l})-\phi'(\theta_j)))\ \left[h_{\tilde{p}(X_{j+l+1})}(\theta_{j+l}) - h(\theta_{j+l})\right] \\ &= \gamma_{j+1}\phi'(\theta_j)\overbrace{\sum_{l=0}^{p-1}\left[h_{\tilde{p}(X_{j+l+1})}(\theta_j) - h(\theta_j)\right]}^{0} \\ &\quad +\gamma_{j+1}\phi'(\theta_j)\sum_{l=0}^{p-1}\left(\left[h_{\tilde{p}(X_{j+l+1})}(\theta_{j+l} - h_{\tilde{p}(X_{j+l+1})}(\theta_j)\right] - [h(\theta_{j+l}) - h(\theta_j)]\right) \\ &\quad +\phi'(\theta_j)\sum_{l=0}^{p-1}(\gamma_{j+l+1}-\gamma_{j+1})\left[h_{\tilde{p}(X_{j+l+1})}(\theta_{j+l}) - h(\theta_{j+l})\right] \\ &\quad +\sum_{l=0}^{p-1}\gamma_{j+l+1}[\phi'(\theta_{j+l})-\phi'(\theta_j)]\left[h_{\tilde{p}(X_{j+l+1})}(\theta_{j+l}) - h(\theta_{j+l})\right]\end{aligned}$$

Now since

$$\theta_{k+1} - \theta_k = \gamma_{k+1} H(\theta_k, X_{k+1}) + \gamma_{k+1}^2 \rho_{k+1}(\theta_k, X_{k+1})$$

we have for $l = 1, \cdots, p-1$ (assume $p \geq 2$)

$$\begin{aligned} |\theta_{j+l} - \theta_j| &\leq \sum_{k=0}^{l-1} \left(|\gamma_{j+k+1} H(\theta_{j+k}, X_{j+k+1})| + \left| \gamma_{j+k+1}^2 \rho_{j+k+1}(\theta_{j+k}, X_{j+k+1}) \right| \right) \\ &\leq 2 \sum_{k=0}^{l-1} [C_1(Q) + \tilde{\gamma} C_2(Q)] \gamma_{j+k+1} (1 + |X_{j+k+1}|^{(q_1 \vee q_2)}) \\ &\leq 2 \sum_{k=0}^{p-2} [C_1(Q) + \tilde{\gamma} C_2(Q)] \gamma_{j+k+1} (1 + |X_{j+k+1}|^{(q_1 \vee q_2)}) \end{aligned}$$

Hence we have (for $p \geq 2$)

$$\begin{aligned} \left| \sum_{l=0}^{p-1} \varepsilon_{j+l}^{(5)} \right| &\leq 2\gamma_{j+1} M_1(Q) \tilde{L}(Q) \sum_{l=0}^{p-1} |\theta_{j+l} - \theta_j| \\ &\quad + 2M_1(Q) \tilde{M}(Q) \sum_{l=0}^{p-1} |\gamma_{j+l+1} - \gamma_{j+1}| \\ &\quad + 2M_2 \tilde{M}(Q) \sum_{l=0}^{p-1} \gamma_{j+l+1} |\theta_{j+l} - \theta_j| \\ &\leq 4(p-1) M_1(Q) \tilde{L}(Q) [C_1(Q) + \tilde{\gamma} C_2(Q)] \sum_{k=0}^{p-2} \tilde{\gamma}_j^2 (1 + |X_{j+k+1}|^{(q_1 \vee q_2)}) \\ &\quad + 2(p-1) M_1(Q) \tilde{M}(Q) \sum_{l=1}^{p-1} |\gamma_{j+l+1} - \gamma_{j+l}| \\ &\quad + 4(p-1) M_2 \tilde{M}(Q) [C_1(Q) + \tilde{\gamma} C_2(Q)] \sum_{k=0}^{p-2} \tilde{\gamma}_j^2 (1 + |X_{j+k+1}|^{(q_1 \vee q_2)}) \end{aligned}$$

Here $\tilde{\gamma}_j = \max_{l=1,\ldots,p-1} \gamma_{j+l}$. Let

$$K_5 = 4(p-1)[C_1(Q) + \tilde{\gamma} C_2(Q)](M_1(Q) \tilde{L}(Q) + M_2 \tilde{M}(Q))$$

Then we have for $0 \leq n \leq m \wedge \tau - 1$,

$$\left| \sum_{k=0}^{n} \varepsilon_k^{(5)} \right| \leq I_{\{\lfloor \frac{n+1}{p} \rfloor \geq 1\}} \left(\sum_{k=0}^{\lfloor \frac{n+1}{p} \rfloor - 1} K_5 \sum_{l=0}^{p-2} \tilde{\gamma}_{kp}^2 (1 + |X_{kp+l+1}|^{(q_1 \vee q_2)}) \right.$$

$$+ 2(p-1)M_1(Q)\tilde{M}(Q) \underbrace{\sum_{k=0}^{\lfloor \frac{n+1}{p} \rfloor -1} \sum_{l=1}^{p-1} |\gamma_{kp+l+1} - \gamma_{kp+l}|}_{\leq \tilde{K}_\gamma} \Bigg)$$

$$+ I_{\{\lfloor \frac{n+1}{p} \rfloor p \neq (n+1)\}} \underbrace{\sum_{k=\lfloor \frac{n+1}{p} \rfloor p}^{n} \gamma_{k+1} 2M_1(Q)\tilde{M}(Q)}_{\leq 2(p-1)M_1(Q)\tilde{M}(Q)\tilde{\gamma}}$$

Remark 3.3 *Note that the last term actually tends to zero if $\gamma_k \to 0$ and $n \to \infty$.*

□

Also for $0 \leq n \leq m \wedge \tau - 1$

$$\begin{aligned}
&\left(I_{\{\lfloor \frac{n+1}{p} \rfloor p \neq (n+1)\}} \sum_{k=\lfloor \frac{n+1}{p} \rfloor p}^{n} \gamma_{k+1} 2M_1(Q)\tilde{M}(Q) \right)^2 \\
&\leq \; I_{\{\lfloor \frac{n+1}{p} \rfloor p \neq (n+1)\}} 4M_1^2(Q)\tilde{M}^2(Q)(p-1) \sum_{k=\lfloor \frac{n+1}{p} \rfloor p}^{n} \gamma_{k+1}^2 \\
&\leq \; I_{\{\lfloor \frac{n+1}{p} \rfloor p \neq (n+1)\}} 4M_1^2(Q)\tilde{M}^2(Q)(p-1)^2 \sum_{k=0}^{m-1} \gamma_{k+1}^2
\end{aligned}$$

Let

$$K_6 = 2(p-1)M_1(Q)\tilde{M}(Q)$$

Thus we have for $p \geq 2$,

$$\begin{aligned}
&E_{x,a}\{\sup_{n \leq m} I(n \leq \tau) \left| \sum_{k=0}^{n-1} \varepsilon_k^{(5)} \right|^2 \} \\
&\leq \; 4K_6^2[\tilde{K}_\gamma^2 + \sum_{k=0}^{m-1} \gamma_{k+1}^2] \\
&\quad + 2K_5^2 E_{x,a} \left\{ I_{\{\lfloor \frac{m}{p} \rfloor \geq 1\}} \left(\sum_{k=0}^{\lfloor \frac{m}{p} \rfloor -1} \sum_{l=0}^{p-2} \tilde{\gamma}_{kp}^2 (1 + |X_{kp+l+1}|^{(q_1 \vee q_2)}) I(kp+l+1 \leq \tau) \right) \right\}^2
\end{aligned}$$

$$
\begin{aligned}
&\leq 4K_6^2[\tilde{K}_\gamma^2 + \sum_{k=0}^{m-1}\gamma_{k+1}^2] + 2K_5^2 \overbrace{\left(I_{\{\lfloor \frac{m}{p}\rfloor \geq 1\}} \sum_{k=0}^{\lfloor \frac{m}{p}\rfloor -1}\sum_{l=0}^{p-2}\tilde{\gamma}_{kp}^2\right)}^{\leq (p-1)\sum_{k=0}^{m-1}\gamma_{k+1}^2} \\
&\quad \cdot E_{x,a}\left\{I_{\{\lfloor \frac{m}{p}\rfloor \geq 1\}} \sum_{k=0}^{\lfloor \frac{m}{p}\rfloor -1}\sum_{l=0}^{p-2}\tilde{\gamma}_{kp}^2(1+|X_{kp+l+1}|^{(q_1\vee q_2)})^2 I(kp+l+1\leq\tau)\right\} \\
&\leq 4K_6^2[\tilde{K}_\gamma^2 + \sum_{k=0}^{m-1}\gamma_{k+1}^2] \\
&\quad +4K_5^2(p-1)^2\mu_{2(q_1\vee q_2)}(Q)\left(\sum_{k=0}^{m-1}\gamma_{k+1}^2\right)^2(1+|x|^{2(q_1\vee q_2)})
\end{aligned}
$$

The first inequality comes from the fact that for any positive integer n and real numbers a_i we have $\left(\sum_{i=1}^n a_i\right)^2 \leq n(\sum_{i=1}^n a_i^2)$ by Schwartz inequality. The second inequality essentially comes from Schwartz inequality.

Moreover on $\{\tau = +\infty\}$, $\sum_{k=0}^{n-1}\varepsilon_k^{(5)}$ converges a.s. and in L^2 if $\sum\gamma_{k+1}^2 < \infty$. See Remark 3.3. Also note that for real numbers a and b, $2ab \leq (a^2+b^2)$.

Combining the above results we have the following, which is the equivalent of Proposition 7 in [8, pages 228–229].

□

Proposition 3.1 *Assume $\tilde{\gamma} \leq 1$. For any compact subset Q of D, and for any C^2 function ϕ on $\mathbf{R}^d$ with bounded second derivatives, there exist constants B_1, B_2 and s such that for all $m \geq 1$:*

1. *We have*

$$
\begin{aligned}
&E_{x,a}\left\{\sup_{n\leq m} I(n\leq\tau(Q))\left|\sum_{k=0}^{n-1}\varepsilon_k(\phi)\right|\right\}^2 \leq \\
&\qquad B_1(1+|x|^s)(1+\sum_{k=0}^{m-1}\gamma_{k+1}^{2\lambda})(\tilde{K}_\gamma^2+\sum_{k=0}^{m-1}\gamma_{k+1}^2) \qquad (3.27)
\end{aligned}
$$

where λ is the constant $\in [1/2, 1]$ of (A.4); and similarly making explicit the constants of Assumptions (A.3), (A.4) and (A.5)

$$\begin{aligned} B_1 \;\leq\; & \tilde{B}_1(1+\mu_s^3(Q)) \\ & \Big[M_1^4(Q) + C_1^4(Q) + C_2^4(Q) + C_3^4(Q) \\ & + \tilde{M}^4(Q) + \bar{C}^{4\lambda}(Q) C_4^4(Q) + M_1^4(Q)\tilde{L}^4(Q)\Big] \end{aligned}$$

where $\bar{C}(Q) = C_1(Q) + \tilde{\gamma} C_2(Q)$, $\tilde{B}_1$ being independent of Q. Lastly we may take $s = \max(2q_4 + 2\lambda(q_1 \vee q_2), 2q_3 + 2(q_1 \vee q_2), 4q_1, 4q_2)$.

2. *If $\sum_{k\geq 1} \gamma_k^{1+\lambda} < \infty$*

   (i)

$$\begin{aligned} & E_{x,a}\left\{\sup_n I(n \leq \tau(Q)) \left|\sum_{k=0}^{n-1} \varepsilon_k(\phi)\right|\right\}^2 \\ & \qquad \leq \; B_2(1+|x|^s)(\tilde{K}_\gamma^2 + \sum_{k\geq 1} \gamma_k^{1+\lambda}) \end{aligned} \tag{3.28}$$

   where $B_2 \leq C B_1$ for some constant C independent of Q but depending on the sequence $\{\gamma_k\}$. In particular $C \leq \tilde{C}(1+\sum_{k=0}^{\infty} \gamma_{k+1}^{1+\lambda})$ for some constant $\tilde{C}$.

   (ii) *On $\{\tau(Q) = \infty\}$ the series $\sum_k \varepsilon_k(\phi)$ converges a.s. and in L^2.*

□

Proof of Proposition 3.1

Essentially the same as the proof for Proposition 7 in [8, pages 228–229], taking into consideration the additional bounds given for

$$E_{x,a}\{\sup_{n\leq m} I(n \leq \tau) \left|\sum_{k=0}^{n-1} \varepsilon_k^{(5)}\right|^2\}$$

by Lemma 3.7.

□

The restriction that $\tilde{\gamma} \leq 1$ can be removed if we allow $\tilde{B}_1$ to be dependent on $\tilde{\gamma}$.

Corollary 3.1 *For all* $T > 0$

$$E_{x,a}\left\{\sup_{n\leq m(T)} I(n \leq \tau(Q))|\sum_{k=0}^{n-1}\varepsilon_k(\phi)|\right\}^2$$
$$\leq \quad B_1(1+|x|^s)(1+T\tilde{\gamma}^{2\lambda-1})(\tilde{K}_\gamma^2 + \sum_{k=1}^{m(T)}\gamma_k^2) \tag{3.29}$$

□

The assumption $\tilde{\gamma} \leq 1$, is introduced to simplify the expression of the constants. It is unimportant, since it can always be obtained by modifying H and ρ_n.

Let $P_{n,x,a}$ denote the distribution of (X_{n+k}, θ_{n+k}) with $X_n = x,\ \theta_n = a$. We introduce the following assumptions [8, page 233].

(A.6) $\sum_{n\geq 1}\gamma_n^{1+\lambda} < +\infty$, where λ is given by (A.4-iii).

□

(A.7) There exists a positive function U of class C^2 on D such that $U(\theta) \to C \leq +\infty$ if $\theta \to \partial D$ or $|\theta| \to +\infty$ and $U(\theta) < C$ for $\theta \in D$ satisfying:

$$U'(\theta)\cdot h(\theta) \leq 0 \text{ for all } \theta \in D$$

□

Let F be a compact subset of D satisfying for some non-negative real number c_0,

$$F = \{\theta : U(\theta) \leq c_0\} \supset \{\theta : U'(\theta)\cdot h(\theta) = 0\} \tag{3.30}$$

We add the following remarks.

Remark 3.4 *Essentially the result of Theorem 9 of [8, page 232] holds but with the modification brought about by replacing Proposition 7 in [8, page 228-229] by Proposition 3.1. The results given by Proposition 10, Proposition 11, Lemma 12, Theorem 13 and Theorem 15 of [8, pages 234–238] hold, with $\tilde{\gamma}$ replacing γ_1 wherever it appears. (Note the changes in the values of the constants brought about by replacing Proposition 7 in [8, pages 228–229] by Proposition 3.1).*

□

In particular we restate Theorem 13 of [8, page 236] with the proper modifications.

Theorem 3.1 *We assume that (A.1) to (A.7) hold and that F is a compact set satisfying 3.30. Then for any compact $Q \subset D$, there exist constants B_4 and s such that for all $n \geq 0$, all $a \in Q$, all x,*

$$P_{n,x,a}\{\theta_k \text{ converges to } F\} \geq 1 - B_4(1 + |x|^s)\left((K_\gamma^{n+1})^2 + \sum_{k=n+1}^{+\infty} \gamma_k^{1+\lambda}\right)$$

□

We have the following global convergence theorem, which is essentially the same as Theorem 17 of [8, page 239] but under our modified assumptions.

3.4 A Convergence Theorem

3.4.1 Assumptions

Assume that the constants $C_i(Q)$ of Assumptions (A.3) and (A.4) grow atmost linearly with the diameter of Q, the constant C_4 being independent of Q if $\lambda = 1$

and the order of $(\mathrm{diam}(Q))^{1-\lambda}$ if $\lambda < 1$. Also we suppose that the constants μ_q in Assumption (A.5) are independent of Q. We make the following *additional assumption* that $\tilde{M}(Q)$, the bound on the magnitude of $h_i(\theta), i = 0, \cdots, p-1$ grows atmost linearly with the diameter of Q and the Lipschitz constant for $h_i(\theta)$ is independent of Q. Thus we suppose the existence of constants $\bar{C}_i,\ q_i,\ i = 1, \ldots, 4,\ \bar{M}, \bar{L}$ and μ_q $(q > 0)$, such that for all $\theta \in \mathbf{R}^d, a \in \mathbf{R}^d, n \geq 0, R > 0$, we have:

$$
\begin{aligned}
|H(\theta, x)| &\leq \bar{C}_1(1 + |\theta|)(1 + |x|^{q_1}) && (3.31)\\
|\rho_n(\theta, x)| &\leq \bar{C}_2(1 + |\theta|)(1 + |x|^{q_2}) && (3.32)\\
E_{x,a}\{1 + |X_{n+1}|^q\} &\leq \mu_q(1 + |x|^q) && (3.33)\\
|\nu_\theta(x)| &\leq \bar{C}_3(1 + |\theta|)(1 + |x|^{q_3}) && (3.34)\\
|h_i(\theta)| &\leq \bar{M}(1 + |\theta|) \qquad i = 0, \cdots, p-1 && (3.35)\\
|h(\theta)| &\leq \bar{M}(1 + |\theta|) && (3.36)
\end{aligned}
$$

and for all θ, θ' such that $|\theta| \leq R,\ |\theta'| \leq R$ and some $\lambda \in [\frac{1}{2}, 1]$,

$$
\begin{aligned}
&|\Pi_\theta \nu_\theta(x) - \Pi_{\theta'} \nu_{\theta'}(x)| \\
&\leq \bar{C}_4(1 + R^{1-\lambda})|\theta - \theta'|^\lambda (1 + |x|^{q_4}) && (3.37)
\end{aligned}
$$

where ν_θ satisfies Assumption (A.4-ii). Also for all θ, θ'

$$
\begin{aligned}
|h_i(\theta) - h_i(\theta')| &\leq \bar{L}|\theta - \theta'| && (3.38)\\
|h(\theta) - h(\theta')| &\leq \bar{L}|\theta - \theta'| && (3.39)
\end{aligned}
$$

We further assume the existence of a constant $\tilde{K}_\gamma$ and hence that of a constant

$\tilde{\gamma}$ such that

$$\sum_{k=1}^{\infty} |\gamma_k - \gamma_{k+1}| \leq \tilde{K}_\gamma \tag{3.40}$$

$$\sup_k \gamma_k = \tilde{\gamma} \tag{3.41}$$

and

$$\sum_k \gamma_k^{1+\lambda} < +\infty \tag{3.42}$$

3.4.2

Theorem 3.2 *We suppose that Assumptions (A.1), (A.2) and (3.31) to (3.42) are satisfied. Then the following holds:*

a. *if there exists a positive function U on $\mathbf{R}^d$ of class C^2(i.e. continuous second partials exist) with bounded second derivatives such that for all θ, $|\theta| \geq \rho_0$*

(i)

$$U'(\theta) \cdot h(\theta) \leq 0$$

(ii)

$$U(\theta) \geq \alpha|\theta|^2, \qquad \alpha > 0$$

then for all $a \in \mathbf{R}^d, x \in \mathbf{R}^k$, the sequence (θ_n) is $P_{x,a}$ a.s. bounded;

b. *if further there exists $\theta_* \in \mathbf{R}^d$ such that*

(i)'

$$U'(\theta) \cdot h(\theta) < 0 \quad \textit{for all } \theta \neq \theta_*$$

(iii)

$$U(\theta) \;=\; 0 \quad \textit{iff} \quad \theta = \theta_*$$

then the sequence (θ_n) *converges* $P_{x,a}$ *a.s. to* θ_*.

□

Proof of Theorem 3.2

The proof is essentially the same as the proof of Theorem 17 in [8, pages 240– 243] but using Proposition 3.1 instead of Proposition 7 of [8, pages 228–229] in the proof of Lemma 21 of [8, pages 242–243]. Note that Lemmas 18, 19 and 20 of [8, pages 241-242] continue to hold. Also refer to the Remark 3.4. The only other thing to be verified is whether the terms [8, page 243] $B_1(Q_n)$ and $B_2(Q_n)$ can be bounded above by a constant times 2^{2n}, under our modified assumptions, where $Q_n = \{\theta : U(\theta) \leq A2^{n+1}\}$, with A being a constant (as defined in [8, page 240]). But this is indeed true since $\tilde{M}(Q_n)$ increases as a constant times $2^{n/2}$ and $\tilde{L}(Q_n) \leq \bar{L}$.

□

Note that for the algorithm in equation 3.1 to converge we need not know the exact period p of the underlying Markov process (X_n). Let $n > 0$ and focus on a particular sample path (X_n). Then we say that the weightage given to class $m \in \{0, \cdots, p-1\}$ on the discrete time inteval $[n, n+N]$ for some $N > 0$ is

$$W_{n,n+N}(m) \equiv \frac{\sum_{k=n}^{n+N} \gamma_k I_{\{\tilde{p}(X_k)=m\}}}{\sum_{k=n}^{n+N} \gamma_k}$$

noting the fact that $\tilde{p}(X_{k+1}) = (\tilde{p}(X_k) + 1) \bmod p$. We define the weight vector W

with $W(m) = \lim_{N\to\infty} W_{1,N}(m)$ whenever it exists. For any p dimensional vector w (with $w(i) \geq 0, \sum_{i=0}^{p-1} w(i) = 1$), define $h^w(\theta) \equiv \sum_{i=0}^{p-1} w(i)h_i(\theta)$. If the weight vector W exists and $W(m) \neq \frac{1}{p}$ for atleast one $m \in \{0, \cdots p-1\}$, then the different periodic classes (when $p > 1$) are being sampled with 'unequal weightage' and hence there is a possibility for the algorithm converging to the point $\tilde{\theta}$ such that $h^W(\tilde{\theta}) = 0$ instead of converging to θ_* such that $h(\theta_*) = 0$. When the condition $\sum_k |\gamma_{k+1} - \gamma_k| < +\infty$ is imposed we have $W(m) = \frac{1}{p}, \forall m \in \{0, \cdots p-1\}$. This ensures that the p different classes are 'sampled' with equal 'weightage'. Also note that for any non-increasing sequence of non-negative numbers (γ_k), $\sum_{k\geq 1} |\gamma_k - \gamma_{k+1}|$ is bounded.

We close this chapter with the following remark.

Remark 3.5 *We may suppose the non-negative step-sizes γ_n to be random. Let γ_{n+1} be measurable w.r.t the sigma-field $\mathcal{F}_n$ (c.f. the proof of Lemma 2 of [8, page 224]) but with the additional restriction that $\sum_k \gamma_k = +\infty$ $P_{x,a}$ a.s. Let there be a deterministic non-negative sequence $\hat{\gamma}_n$, such that $\sum_{k=1}^{\infty} \hat{\gamma}_n^{1+\lambda} < +\infty$, and deterministic non-negative sequence $\hat{\delta}_n$ such that $\sum_{k=1}^{\infty} \hat{\delta}_n < +\infty$ such that $\gamma_n \leq \hat{\gamma}_n$ and $|\gamma_{n+1} - \gamma_n| \leq \hat{\delta}_n$ for all but a finite number of n, $P_{x;a}$ a.s. Then the conclusion of Theorem 3.2 continues to hold. An outline of the proof is as follows. Let*

$$A_n \equiv \{\omega \in \Omega : \gamma_k(\omega) \leq \hat{\gamma}_k, |\gamma_{k+1}(\omega) - \gamma_k(\omega)| \leq \hat{\delta}_k \text{ for all } k \geq n\}$$

. Then A_n is an 'increasing' sequence of sets and that $P_{x,a}(\bigcup_n A_n) = 1$.

□

As a final note see that our result does not deal with the case when the period or periodic classes of the Markov process changes with θ.

Chapter 4

Temporal Difference Schemes For Discounted Cost MDPs

In this chapter we propose a reinforcement learning scheme for finding optimal and sub-optimal policies for the finite state, finite action Markov Decision Problem (MDP) with the infinite horizon discounted cost criterion. Online learning is utilized along with temporal difference schemes for approximating value functions to obtain a direct adaptive control scheme for the MDP. The approach features the approximation of stationary deterministic policies with stationary randomized policies. We provide convergence results of the algorithm under very reasonable assumptions, in particular without aperiodicity assumptions.

In Section 4.2 we discuss Stationary Randomized Policies. Section 4.3 deals with approximate policy iteration [16]. This is followed by Temporal Difference (TD) schemes [16, 26, 50, 54] for estimating the value function (with linear function approximation) for discounted cost Markov Cost processes in Section 4.4. In Section 4.5 we discuss on-line learning schemes for finding optimal and sub-optimal policies for discounted cost MDPs which uses TD schemes for policy evaluation.

4.1 Markov Decision Process Model Revisited

We restate the MDP model discussed in Chapter 1, for convenience. Let $\mathbf{N}$ denote the set of positive integers and $\mathbf{N}_0$ denote the set of non-negative integers. For

a set $\mathcal{A}$, $|\mathcal{A}|$ denotes the cardinality of $\mathcal{A}$, whereas for a real number α, $|\alpha|$ denotes the absolute value of α. Let the non-empty state space of the MDP [4, 12, 23] be $\mathcal{S} = \{1, 2, \cdots, n\}$ and the non-empty control constraint sets be $\mathcal{A}(i) = \{1, 2, \cdots, |\mathcal{A}(i)|\}$, for each $i \in \mathcal{S}$, which denote the possible control actions (feasible actions) from state i. Define the action space $\mathcal{A} = \bigcup_{i=1}^{n} \mathcal{A}(i)$. Note that we are dealing with finite state finite action homogeneous MDPs [11, 12]. The state at time $t \in \mathrm{N}_0$ is denoted by s_t and the action taken at time t is denoted by u_t. The transition probabilities may be conveniently denoted by $p_{ij}(u) = \Pr[s_{t+1} = j \mid s_t = i, u_t = u]$, where $i, j \in \mathcal{S}$ and $u \in \mathcal{A}(i)$. Here Pr denotes probability. $g_t \in \mathbf{R}$ denotes the immediate cost incurred at time t when action $u_t \in \mathcal{A}(s_t)$ is taken from state s_t. The distribution of the immediate cost which may be random is independent of past states, actions and immediate costs, given current state s_t, current action u_t and successive state s_{t+1}. For $u \in \mathcal{A}(i)$ let $g(i, u, j)$ denote the expected value $\mathrm{E}[g_t \mid s_t = i, u_t = u, s_{t+1} = j]$. For $u \in \mathcal{A}(i)$ let $g(i, u)$ denote the expected value $\mathrm{E}[g_t \mid s_t = i, u_t = u]$. Now $g(i, u) = \sum_{j=1}^{n} p_{ij}(u)\, g(i, u, j)$. We assume these expectations to be finite.

The MDP evolves as follows. At time $t = 0$, let the initial state be s_0. If at time $t \in \mathbf{N}_0$, the state is s_t and the control $u_t \in \mathcal{A}(s_t)$ is applied, then an immediate cost g_t (which may be random) is incurred and the system moves to the state s_{t+1} according to the transition probability $p_{s_t, s_{t+1}}(u_t)$. A realization of the process looks like $(s_0, u_0, g_0, s_1, u_1, g_1, \cdots) \in \Omega \equiv (\mathcal{S}\mathcal{A}\mathbf{R})^{\infty}$, where $\mathbf{R}$ is the set of real numbers. $\{u_t\}$ is a control sequence in $\mathcal{A}$ determined by a control policy. $\mathcal{S}$ and $\mathcal{A}$ are endowed with the discrete topology. $\mathbf{R}$ is endowed with the Borel topology. Let $h_t = (s_0, u_0, g_0, s_1, u_1, g_1, \cdots, s_{t-1}, u_{t-1}, g_{t-1}, s_t)$ denote the history

of the process upto time t with $h_0 = (s_0)$. The history follows the recursion $h_t = (h_{t-1}, u_{t-1}, g_{t-1}, s_t)$ for $t \geq 1$. Let $\mathcal{H}_t$ denote the set of all histories upto time t. Here $\mathcal{H}_0 = \mathcal{S}$, $\mathcal{H}_t = \mathcal{H}_{t-1}\mathcal{A}\mathbf{R}\mathcal{S}$. These spaces are endowed with the product topologies. Here $\Omega = \mathcal{H}^\infty = (\mathcal{S}\mathcal{A}\mathbf{R})^\infty$ is the sample space under consideration.

An admissible policy for the MDP is a sequence $\nu = \{\nu_t\}$ such that for each $t \in \mathbf{N}_0$, ν_t is a stochastic kernel on $\mathcal{A}$ given h_t with all the probability measure concentrated on $\mathcal{A}(s_t)$. The set of all admissible policies is denoted by $\mathcal{M}$. The set of all stationary deterministic policies (or control functions to be precise) is denoted by Υ, and the set of all stationary randomized policies (stochastic control kernels to be precise) is denoted by Λ.

A policy $\nu \in \mathcal{M}$ and an initial state s_0, together with the transition probabilities of the MDP and the immediate cost (which may be random for any particular state and action) generating mechanism, determine a unique probability measure denoted by $\mathcal{P}^\nu_{s_0}(\cdot) \equiv \mathcal{P}^\nu(\cdot \mid s_0)$ on the space Ω of all possible realizations of the system [4, 23]. The expectation with respect to this probability is denoted by $\mathrm{E}^\nu[\cdot \mid s_0]$. The performance criterion for the infinite horizon discounted cost problem is the well defined quantity $J^\nu(i) = \mathrm{E}^\nu[\sum_{t=0}^\infty \beta^t g_t \mid s_0 = i]$, the expected total discounted cost when the policy $\nu \in \mathcal{M}$ is used and the initial state is $s_0 = i$. Here $\beta \in [0, 1)$ is the discount factor. The aim is to find a policy $\nu^* \in \mathcal{M}$ such that $J^{\nu^*}(i) = J^*(i)$, $\forall i \in \mathcal{S}$. Here $J^*(i) = \inf_{\nu \in \mathcal{M}} J^\nu(i)$ for $i \in \mathcal{S}$. J^* is called the optimal cost function. It is well known that there exists a stationary deterministic policy [12, 23] which is optimal.

Let $\mathcal{Q} = \{(i, u) | i \in \mathcal{S}, u \in \mathcal{A}(i)\}$, be the set of all state-action pairs.

4.2 Stationary Randomized Policies

Define for each positive integer k, $\Delta_k \equiv \left\{(p_1, p_2, \cdots, p_k) \mid p_l \geq 0; \sum_{l=1}^{k} p_l = 1\right\}$, the $k-1$ dimensional unit simplex. A stationary randomized policy (stochastic control kernel to be precise) can be specified as

$$\delta \in \Lambda$$

where

$$\Lambda \equiv \Delta_{|\mathcal{A}(1)|} \times \Delta_{|\mathcal{A}(2)|} \times \cdots \times \Delta_{|\mathcal{A}(n)|}$$

and for each $i \in \mathcal{S}$

$$\delta(i) \in \Delta_{|\mathcal{A}(i)|}, \quad \text{denotes}$$

$$[\delta(i)]_a = \Pr(a \mid i) = \Pr[u_t = a \mid s_t = i]; \quad a \in \mathcal{A}(i)$$

Here 'Pr' denotes probability. For a particular stationary randomized policy δ, we obtain a homogeneous Markov chain with state space $\mathcal{S}$, transition probability from state i to state j, given by $p_{ij}^{\delta} = \sum_{a \in \mathcal{A}(i)} [\delta(i)]_a \; p_{ij}(a)$ and expected immediate cost from state i given by $\bar{g}^{\delta}(i) = \sum_{a \in \mathcal{A}(i)} [\delta(i)]_a \; g(i,a)$. For $\delta \in \Lambda$, let $P_{\delta} = [p_{ij}^{\delta}]$, denote the corresponding transition probability matrix and $\bar{g}^{\delta} \in \mathbf{R}^n$ denote the expected immediate cost vector whose i^{th} component $\bar{g}^{\delta}(i)$ is the expected immediate cost from state i under policy δ. A stationary deterministic policy or equivalently a control function $\mu \in \Upsilon$ may be regarded as a special case of stationary randomized policy (or stochastic control kernel) in which the probability distribution on the set of actions is degenerate, i.e. all the probability is concentrated on one action, namely $\mu(i)$ for each $i \in \mathcal{S}$. The infinite horizon discounted cost function for the

policy $\delta \in \Lambda$, denoted by $J^\delta \in \mathbf{R}^n$ is given by

$$J^\delta(i) = \mathrm{E}^\delta \left[\sum_{t=0}^{\infty} \beta^t g_t \mid s_0 = i \right]$$

Actually $J^\delta = (I - \beta P_\delta)^{-1} \bar{g}^\delta$. Equivalently we may think of another Markov Chain with state space

$$\mathcal{Q} = \{(i, a) \mid i \in \mathcal{S}, a \in \mathcal{A}(i)\}$$

and transition probability

$$p^\delta_{(i,a)(j,b)} = p_{ij}(a)[\delta(j)]_b$$

The expected immediate cost from "state" (i, a) is given by $g(i, a)$. Then

$$V^\delta(i, a) = \mathrm{E}^\delta \left[\sum_{t=0}^{\infty} \beta^t g_t \mid (s_0, u_0) = (i, a) \right]$$

represents the expected discounted cost of starting from "state" (i, a) for this new Markov Chain. It is easy to see that

$$J^\delta(i) = \sum_{a \in \mathcal{A}(i)} [\delta(i)]_a \; V^\delta(i, a)$$

We introduce the following function $\mathsf{h} : (i, a, V) \mapsto \mathbf{R}$ as follows

$$\mathsf{h}(i, a, V) = g(i, a) + \beta \sum_{j \in \mathcal{S}} p_{ij}(a) \; V(j)$$

for each $i \in \mathcal{S},\ a \in \mathcal{A}(i),\ V \in \mathbf{R}^n$.

For $J \in \mathbf{R}^n$, let $T_\delta : \mathbf{R}^n \to \mathbf{R}^n$ be

$$\begin{aligned} (T_\delta J)(i) &= \bar{g}^\delta(i) + \beta \sum_{j=1}^{n} p^\delta_{ij} \; J(j) \\ &= \sum_{a \in \mathcal{A}(i)} [\delta(i)]_a \; \mathsf{h}(i, a, J) \end{aligned}$$

T_δ is a monotone contraction mapping (with contraction coefficient β) with respect to the supremum norm [12, 16]. See also Section 1.1.2.2.

In fact J^δ is the unique fixed point of the contraction mapping T_δ. Thus

$$J^\delta(i) = \sum_{a \in \mathcal{A}(i)} [\delta(i)]_a \underbrace{\left(g(i,a) + \beta \sum_{j=1}^{n} p_{ij}(a) J^\delta(j) \right)}_{Q^\delta(i,a)}$$

It is easy to see from the definition that $Q^\delta(i,a)$ is the expected discounted cost of taking action a from state i at time $t = 0$ and from then on following the policy δ. Note that $V^\delta(i,a) = Q^\delta(i,a)$. $Q^*(i,a) = h(i,a,J^*)$ denotes the optimal Q-values, where $J^* \in \mathbf{R}^n$ is the optimal cost to go vector for the discounted cost MDP. Let $\tilde{\delta}$ be another stationary randomized policy such that

$$\sum_{a \in \mathcal{A}(i)} \left[\tilde{\delta}(i)\right]_a Q^\delta(i,a) \;\leq\; J^\delta(i), \quad \forall i \in \mathcal{S}$$

Then it follows from the monotonicity property [12] of the operator T_δ that $J^{\tilde{\delta}} \leq J^\delta$, the inequality is componentwise. Let $\alpha_i > 0,\ i \in \{1, \ldots, n\}$. Then it follows that any local minimum of $\mathsf{s}(\delta) \equiv \sum_{i=1}^n \alpha_i J^\delta(i)$ is also a global minimum of $\mathsf{s}(\delta)$ in the domain Λ. Denote by

$$\Lambda_{\bar{\epsilon}} \equiv \{\delta \in \Lambda \mid [\delta(i)]_a \geq \bar{\epsilon}(i),\ i \in \mathcal{S}, a \in \mathcal{A}(i)\}$$

where $\bar{\epsilon} \in \mathbf{R}^n$ with $\bar{\epsilon}(i) \geq 0, \quad \forall i \in \mathcal{S}$. Here $\bar{\epsilon}(i)$ denotes the i^{th} component of $\bar{\epsilon}$. Let $\tilde{\epsilon} \in \mathbf{R}^n$ be the vector with $\tilde{\epsilon}(i) = \frac{1}{|\mathcal{A}(i)|}, \quad \forall i \in \mathcal{S}$.

Then $\underline{0} \leq \bar{\epsilon} \leq \tilde{\epsilon}$ implies that $\Lambda_{\bar{\epsilon}}$ is nonempty, where $\underline{0}$ is the vector with all components equal to zero and the inequality is componentwise. Also $\underline{0} \leq \bar{\epsilon} \leq \hat{\epsilon} \leq \tilde{\epsilon}$

implies that $\Lambda_{\hat{\epsilon}} \subset \Lambda_{\bar{\epsilon}}$. For each positive integer k and $0 \leq \epsilon \leq \frac{1}{k}$ define

$$\Delta_k^\epsilon \equiv \left\{ (p_1, \cdots, p_k) \mid p_i \geq \epsilon, \sum_{i=1}^{k} p_i = 1 \right\}$$

We define the k *extremal points* of Δ_k^ϵ (when $0 \leq \epsilon < \frac{1}{k}$) as follows; the i^{th} one is defined as the probability vector $(p_1, p_2, \cdots, p_k)$ with

$$\begin{aligned} p_i &= (1 - (k-1)\epsilon) \\ p_j &= \epsilon; \quad j \neq i \end{aligned}$$

Note that when $k > 1$, $p_i > p_j$ for $j \neq i$. Also $\Lambda_{\underline{0}} = \Lambda$. A $\delta \in \Lambda_{\bar{\epsilon}}$, with $\underline{0} \leq \bar{\epsilon} < \tilde{\epsilon}$ is called an *extremal policy* of $\Lambda_{\bar{\epsilon}}$ if $\delta(i)$ is an extremal point of $\Delta_{|\mathcal{A}(i)|}^{\bar{\epsilon}(i)}$ for each $i \in \mathcal{S}$. The strict inequality holds component wise.

Observe that the extremal policies of $\Lambda_{\underline{0}}$ are precisely the stationary deterministic policies. Let Υ denote the set of stationary deterministic policies (or control functions to be precise). We will use the notation μ exclusively to denote stationary deterministic policies. Note that there is a natural one to one correspondence between the elements of Υ and the extremal policies of $\Lambda_{\bar{\epsilon}}$ when $\underline{0} \leq \bar{\epsilon} < \tilde{\epsilon}$. An extremal policy δ of $\Lambda_{\bar{\epsilon}}$ corresponding to a stationary stationary deterministic policy $\mu \in \Upsilon$ has the property that $[\delta(i)]_{\mu(i)} > [\delta(i)]_a$ if $a \in \mathcal{A}(i),\ a \neq \mu(i)$ for $i \in \{1, 2, \ldots, n\}$. Without loss of generality we will use $\mu \in \Upsilon$ to denote either the extremal policies of $\Lambda_{\underline{0}}$ or the corresponding control law mapping the states in $\mathcal{S}$ to the corresponding action in each state on which all the probability mass is concentrated. It will be clear from the context whether $\mu(i),\ i \in \mathcal{S}$ denotes an extremal point of $\Delta_{|\mathcal{A}(i)|}^0$ or the corresponding action in $\mathcal{A}(i)$.

For any positive integer k and any $w \in \mathbf{R}^k$ let $\| w \|_1$ denote the ℓ_1 norm defined by $\sum_{i=1}^k |w_i|$. Similarly for any $w \in \mathbf{R}^k$ define $\| w \|$ to be the ℓ_∞ or supremum norm, namely $\max_{i \in \{1,\ldots,k\}} |w_i|$. We define a metric d on the set Λ. For any $\delta, \tilde{\delta} \in \Lambda$, define

$$\mathsf{d}(\delta, \tilde{\delta}) \equiv \max_{i \in \mathcal{S}} \ \| \delta(i) - \tilde{\delta}(i) \|_1$$

It is easy to verify that this is a metric and further that Λ is a compact space under this metric. Define

$$\text{Interior}(\Lambda) \ \equiv \ \bigcup_{\bar{\epsilon}: \underline{0} < \bar{\epsilon} \leq \tilde{\epsilon}} \Lambda_{\bar{\epsilon}}$$

Note that δ is an element of Interior(Λ) if and only if δ assigns positive probabilities to each possible action from each state. Such policies are called *stationary fully randomized policies*. Since P_δ and $\bar{g}^\delta$ are continuous functions on the space Λ, it follows that the cost to go vector J^δ is a continuous function on Λ. In fact the compactness of Λ implies that J^δ is uniformly continuous on Λ. In particular given any $\epsilon > 0$, there exists $\varsigma > 0$ (dependent on ϵ) such that $\| J^\mu - J^\delta \| < \epsilon$ for each $\mu \in \Upsilon$ and $\delta \in \Lambda$ with $\mathsf{d}(\mu, \delta) < \varsigma$.

A policy $\mu \in \Upsilon$ is said to be a *greedy policy* for $V \in \mathbf{R}^n$ if

$$\mu(i) = \arg \min_{a \in \mathcal{A}(i)} \mathsf{h}(i, a, V) \qquad \forall i \in \mathcal{S}.$$

or equivalently

$$\mathsf{h}(i, \mu(i), V) = \min_{a \in \mathcal{A}(i)} \mathsf{h}(i, a, V) \qquad \forall i \in \mathcal{S}.$$

Note that the dynamic programming operator $T : \mathbf{R}^n \to \mathbf{R}^n$ is actually given by

$$(TV)(i) = \min_{a \in \mathcal{A}(i)} \mathsf{h}(i, a, V) \qquad \forall i \in \mathcal{S}$$

and is a monotone contraction mapping with contraction coefficient β, under the supremum norm and has as its unique fixed point J^*. Note that for each $i \in \mathcal{S}$ and $a \in \mathcal{A}(i)$ the function $\mathsf{h}(i, a, \cdot)$ is an affine function on the space $\mathbf{R}^n$. Note that for any $\mu \in \Upsilon$ the function $(T_\mu V)(i) = h(i, \mu(i), V)$. We define for each $\mu \in \Upsilon$ the *greedy region* for μ as

$$\mathcal{R}_\mu = \{V \in \mathbf{R}^n \mid \mu \text{ is greedy for } V\}.$$

It is easy to see that $\mathcal{R}_\mu$ is a polyhedron. Also note that $\mathcal{R}_\mu$ may be empty for some μ and that $\mathbf{R}^n = \bigcup_{\mu \in \Upsilon} \mathcal{R}_\mu$. Since a policy $\mu \in \Upsilon$ is optimal if and only if $T_\mu J^\mu = T J^\mu$, a policy $\mu \in \Upsilon$ is optimal if and only if $J^\mu \in \mathcal{R}_\mu$. In fact such optimal $\mu^* \in \Upsilon$ exists [12].

4.3 Approximate Policy Iteration

For $V \in \mathbf{R}^n$, let $\tilde{\mathsf{h}}(i, a, V)$ denote an approximation to $\mathsf{h}(i, a, V)$ for each $i \in \mathcal{S}$ and $a \in \mathcal{A}(i)$. We have the following lemma.

Lemma 4.1 *Let $\tilde{V}$ be any fixed vector in $\mathbf{R}^n$. Then there exist scalars $\epsilon > 0,\ \varsigma > 0$ dependent on $\tilde{V}$ such that if V is any vector in $\mathbf{R}^n$ with $\| V - \tilde{V} \| < \epsilon$ and $\tilde{\mathsf{h}}$ is such that $|\tilde{\mathsf{h}}(i, a, V) - \mathsf{h}(i, a, V)| < \varsigma,\ \forall a \in \mathcal{A}(i),\ i \in \mathcal{S}$; then the control policy $\tilde{\mu} \in \Upsilon$ obtained by setting $\tilde{\mu}(i) = \arg\min_{a \in \mathcal{A}(i)} \tilde{\mathsf{h}}(i, a, V)$ for each $i \in \mathcal{S}$ is a greedy policy for the vector $\tilde{V}$.*

□

Proof of Lemma 4.1

Note that for any $i \in \mathcal{S}$ and $a \in \mathcal{A}(i)$ the operator $\mathsf{h}(i, a, \cdot)$ is an affine function with the property that $\mathsf{h}(i, a, U + \alpha \underline{1}) = \mathsf{h}(i, a, U) + \beta\, \alpha \underline{1}$ for $U \in \mathbf{R}^n,\ \alpha \in \mathbf{R}$. Here

$\underline{1} \in \mathbf{R}^n$ is the vector with all components equal to one. Also $\mathsf{h}(i,a,\cdot)$ is monotone, i.e. if $U, J \in \mathbf{R}^n$ and $U \geq J$ then $\mathsf{h}(i,a,U) \geq \mathsf{h}(i,a,J)$.

Now in the Lemma 4.1, $|\mathsf{h}(i,a,V) - \mathsf{h}(i,a,\tilde{V})| < \beta\epsilon$. We can choose $\epsilon > 0,\ \varsigma > 0$ such that

$$\mathsf{h}(i,u,\tilde{V}) - \min_{a \in \mathcal{A}(i)} \mathsf{h}(i,a,\tilde{V}) > 2\beta\epsilon + 2\varsigma$$

$\forall i \in \mathcal{S},\ u \in \mathcal{A}(i)$ such that $u \neq \arg\min_{a \in \mathcal{A}(i)} \mathsf{h}(i,a,\tilde{V})$. Hence if $\tilde{\mu}(i) = \arg\min_{a \in \mathcal{A}(i)} \tilde{\mathsf{h}}(i,a,V)$ for $i \in \mathcal{A}(i)$, then

$$\mathsf{h}(i,\tilde{\mu}(i),\tilde{V}) = \min_{a \in \mathcal{A}(i)} \mathsf{h}(i,a,\tilde{V}), \qquad \forall i \in \mathcal{S}.$$

□

We have the following corollary to Lemma 4.1.

Corollary 4.1 *For any finite state, finite action MDP there exist scalars* $\epsilon > 0,\ \varsigma > 0$ *such that if* J *is any vector in* $\mathbf{R}^n$ *with* $\| J - J^\mu \| < \epsilon$ *and* $|\tilde{\mathsf{h}}(i,a,J) - h(i,a,J)| < \varsigma,\ \forall a \in \mathcal{A}(i),\ i \in \mathcal{S}$*; then the control policy* $\tilde{\mu}$ *obtained by setting* $\tilde{\mu}(i) = \arg\min_{a \in \mathcal{A}(i)} \tilde{\mathsf{h}}(i,a,J)$ *for each* $i \in \mathcal{S}$ *is a greedy policy for the vector* J^μ*. In fact the* ϵ *and* ς *are uniformly applicable to all* $\mu \in \Upsilon$*.*

□

Proof of Corollary 4.1

The proof follows from Lemma 4.1 and the fact that $|\Upsilon|$ is finite.

□

Assume that a sequence of stationary deterministic policies μ_k and a corresponding sequence of approximate cost-to-go functions J_k satisfy

$$\max_{i \in \mathcal{S}} |J_k(i) - J^{\mu_k}(i)| \ \leq \ \epsilon, \quad \text{for } k = 0, 1, \ldots \tag{4.1}$$

and

$$\max_{i \in \mathcal{S}} |(T_{\mu_{k+1}} J_k)(i) - (T J_k)(i)| \ \leq \ \varsigma, \quad \text{for } k = 0, 1, \ldots \tag{4.2}$$

where ϵ and ς are some positive scalars. Then we have the following lemma from [16, Proposition 6.2, page 276].

Lemma 4.2 *A sequence of policies μ_k and functions J_k satisfying inequalities (4.1) and (4.2) satisfy*

$$\limsup_{k \to \infty} \| J^{\mu_k} - J^* \| \ \leq \ \frac{\varsigma + 2\beta\epsilon}{(1-\beta)^2}.$$

□

We may use Lemma 4.2 to prove the next result.

Consider the following algorithm. Pick some $\mu_0 \in \Upsilon$. The sequence $\{\mu_k\}$ of stationary deterministic policies is generated as follows. Let $\delta_k \in \Lambda$ be a sequence of stationary randomized policies generated in such a manner that $\| J^{\delta_k} - J^{\mu_k} \| \leq \epsilon_k$. Let $Q^{\delta_k}(i,a) = h(i,a,J^{\delta_k})$. Let $\tilde{Q}_k(i,a), a \in \mathcal{A}(i), i \in \mathcal{S}$, be such that $|\tilde{Q}_k(i,a) - Q^{\delta_k}(i,a)| \leq \varsigma_k, \forall a \in \mathcal{A}(i), i \in \mathcal{S}$. We set

$$\mu_{k+1}(i) = \arg \min_{a \in \mathcal{A}(i)} \tilde{Q}_k(i,a) \qquad \forall i \in \mathcal{S}.$$

Note that

$$\min_{i \in \mathcal{S}} |(T_{\mu_{k+1}} J^{\delta_k})(i) - (T J^{\delta_k})(i)| \leq 2\varsigma_k.$$

Corollary 4.2 *Suppose* $\epsilon = \limsup_{k\to\infty} \epsilon_k$ *and* $\varsigma = \limsup_{k\to\infty} \varsigma_k$. *Then*

$$\limsup_{k\to\infty} \| J^{\mu_k} - J^* \| \leq \ 2 \ \frac{(\varsigma + \beta\epsilon)}{(1-\beta)^2}$$

□

Note that if ς and ϵ are sufficiently small then $J^{\mu_k} = J^*$ for all large k, since Υ has finite cardinality. Observe that by Corollary 4.1, there exists $\epsilon > 0$ and $\varsigma > 0$ such that if $\epsilon_k < \epsilon$ and $\varsigma_k < \varsigma, \ \forall k$ then the μ_ks obtained are the same ones obtained while doing policy iteration and hence converges to optimal policy in a finite ($\leq |\Upsilon|$) number of steps.

4.4 Temporal Difference (TD(λ)) Schemes

Consider a homogeneous Markov Cost process [54] with state space $\mathcal{S} = \{1, 2, \cdots, n\}$, and transition probability matrix $P = [p_{ij}]$. Let g_t denote the immediate cost incurred while making a transition from state i_t to state i_{t+1} at time $t \in \mathbf{N}_0$, the cost may be random but has finite mean and variance. The probability distribution of g_t may depend on states i_t and i_{t+1}, but given i_t and i_{t+1} does not depend on the past values of i_l and g_l ($l < t$). Let $g(i) \equiv \mathrm{E}[g_t \mid i_t = i]$, and $\bar{g} \in \mathbf{R}^n$ denote the expected immediate cost vector with $\bar{g}(i) = g(i)$. . We are interested in obtaining the value function $J : \mathcal{S} \to \mathbf{R}$ given by

$$J(i) \equiv \mathrm{E}\left[\sum_{t=0}^{\infty} \beta^t g_t \mid i_0 = i\right] = (I - \beta P)^{-1} \bar{g}.$$

Here $\beta \in [0, 1)$. Neither P nor the distribution of g is known in advance. In a general setting of the TD(λ) scheme [19, 20, 26, 50, 54] the aim is to approximate

J using $\tilde{J}(\cdot, r) = \sum_{k=1}^{K} r(k)\phi_k$, where $r = (r(1), \cdots, r(K))^T \in \mathbf{R}^K$ is a parameter vector; $\phi_k \in \mathbf{R}^n$, $k = 1, \ldots, K$ are basis functions. Essentially the interest is in finding $r \in \mathbf{R}^K$, such that some error metric between J and $\tilde{J}(\cdot, r)$ is minimized. Define $\phi(i)$ by $\phi(i) = (\phi_1(i), \cdots, \phi_K(i))^T$. With this notation

$$\begin{aligned} \tilde{J}(i, r) &= r'\phi(i) \\ \tilde{J}(r) &= \Phi r \end{aligned}$$

where

$$\Phi = [\phi(1) \mid \phi(2) \mid \cdots \mid \phi(n)]^T \in \mathbf{R}^{n \times K}$$

Note that the k^{th} column of Φ is ϕ_k. See that

$$\nabla \tilde{J}(i, r) = \phi(i),$$

is the gradient vector for $\tilde{J}(i, r)$, and

$$\nabla \tilde{J}(r) = \Phi'$$

is the Jacobian matrix. Define the temporal difference as

$$d_t = g_t + \beta \tilde{J}(i_{t+1}, r_t) - \tilde{J}(i_t, r_t)$$

where r_t is the parameter vector at time t. For $\lambda \in [0, 1]$, TD(λ) updates r_t according to

$$\begin{aligned} r_{t+1} &= r_t + \gamma_t d_t \sum_{k=0}^{t} (\beta\lambda)^{t-k} \nabla \tilde{J}(i_k, r_t) \\ &= r_t + \gamma_t d_t \sum_{k=0}^{t} (\beta\lambda)^{t-k} \phi(i_k) \end{aligned}$$

where γ_t is a sequence of non-negative scalar step sizes. If we define the sequence of eligibility vectors by

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

then the TD(λ) updates are given by

$$\begin{aligned} z_t &= (\beta\lambda) z_{t-1} + \phi(i_t) \\ r_{t+1} &= r_t + \gamma_t d_t z_t \end{aligned}$$

with

$$z_{-1} = \underline{0}$$

We have the following assumptions.

Assumption 4.1 *Let the following hold.*

(a). *The Markov Chain is irreducible with unique invariant distribution* π *(which satisfies* $\pi' P = \pi'$ *with* $\pi(i) > 0$ *for all* $i \in \mathcal{S}$*).*

(b). Φ *has full column rank, i.e.* $\phi_k' s$ *are linearly independent.*

□

Assumption 4.2 *Let the following hold.*

(a). *The non-negative step sizes* $\gamma_t \downarrow 0$ *are pre-determined and satisfy*

$$\sum_{t=0}^{\infty} \gamma_t = \infty; \quad \sum_{t=0}^{\infty} \gamma_t^2 < \infty.$$

(b). *The immediate cost* g_t *has finite moments, i.e.* $E[(|g_t|)^k \mid i_t = i] < \infty;\ \forall i \in \mathcal{S}, \forall k \in \mathbf{N}$.

□

Actually Assumption 4.2(a) may be replaced by (see Chapter 3)

4.2(a') The non-negative step sizes γ_t are pre-determined and satisfy

$$\sum_{t=0}^{\infty} \gamma_t = \infty, \quad \sum_{t=0}^{\infty} \gamma_t^2 < \infty, \quad \text{and} \quad \sum_{t=0}^{\infty} |\gamma_{t+1} - \gamma_t| < \infty.$$

□

For $\lambda \in [0, 1)$ define the operator $T^{(\lambda)} : \mathbf{R}^n \to \mathbf{R}^n$ as

$$T^{(\lambda)} \bar{J} = (1 - \lambda) \sum_{m=0}^{\infty} \lambda^m \left(\sum_{t=0}^{m} \beta^t P^t \bar{g} + \beta^{m+1} P^{m+1} \bar{J} \right)$$

and

$$T^{(1)} \bar{J} = J = (I - \beta P)^{-1} \bar{g}$$

Here $\bar{J} \in \mathbf{R}^n$. Note that for $\lambda \in [0, 1)$

$$\begin{aligned} T^{(\lambda)} \bar{J} &= (1 - \lambda) \sum_{m=0}^{\infty} \lambda^m \sum_{t=0}^{m} \beta^t P^t \bar{g} + (1 - \lambda) \sum_{m=0}^{\infty} \lambda^m \beta^{m+1} P^{m+1} \bar{J} \\ &= \sum_{m=0}^{\infty} \beta^m \lambda^m P^m \bar{g} + (1 - \lambda) \sum_{m=0}^{\infty} \lambda^m \beta^{m+1} P^{m+1} \bar{J} \\ &= (I - \beta \lambda P)^{-1} \bar{g} + P^{(\lambda)} \bar{J} \end{aligned}$$

where

$$P^{(\lambda)} = (1 - \lambda) \sum_{m=0}^{\infty} \lambda^m \beta^{m+1} P^{m+1}$$

Since $\lim_{\lambda \uparrow 1} P^{(\lambda)} = \mathbf{0}$, the zero matrix, we have

$$\lim_{\lambda \uparrow 1} (T^{(\lambda)} \bar{J}) = (I - \beta P)^{-1} \bar{g} = J = T^{(1)} \bar{J}$$

Let Assumption 4.1 hold. Let D be the $n \times n$ diagonal matrix with diagonal entries $(\pi(1), \pi(2), \ldots, \pi(n))$. For any x, $y \in \mathbf{R}^n$, let the inner product be defined as $< x, y >_D = x'Dy$. The corresponding weighted Eucledian norm is

$$\| x \|_D = \sqrt{< x, x >_D}$$

We say two vectors $\bar{J}, \hat{J} \in \mathbf{R}^n$ are D-orthogonal if$\bar{J}^T D\hat{J} = 0$. Define the projection operator

$$\Pi = \Phi(\Phi' D\Phi)^{-1}\Phi' D$$

Note that

$$\Pi\bar{J} = \arg \min_{\hat{J}\in\{\Phi r | r\in\mathbf{R}^K\}} \| \bar{J} - \hat{J} \|_D$$

Note that $(\bar{J} - \Pi\bar{J})$ is D-orthogonal to ϕ_ks for $k = 1, \ldots, K$ and $\Pi\bar{J}$ is unique.

It may be shown that $T^{(\lambda)}$ and $\Pi T^{(\lambda)}$ are contraction mappings [12, 54] with respect to the weighted Eucledian norm $\| \cdot \|_D$, and has contraction factor [12, Proposition 6.3.3, page 350]

$$\beta_\lambda = \frac{\beta(1-\lambda)}{1-\beta\lambda}$$

In fact $\beta_\lambda = 1 - \frac{(1-\beta)}{1-\beta\lambda}$ and hence $\beta_\lambda \downarrow 0$ as $\lambda \uparrow 1$. Let Φr^* be the unique fixed point of $\Pi T^{(\lambda)}$. The unique fixed point of $T^{(\lambda)}$ is $J = (I - \beta P)^{-1}\bar{g}$.

We have the following result which is an extension of the result in [12, 54], in that the Markov chain need not be aperiodic.

Lemma 4.3 *Under Assumption 4.1 and Assumption 4.2, for any fixed $\lambda \in [0, 1]$, the TD(λ) algorithm converges w.p.1 to a unique $r^* \in \mathbf{R}^K$ irrespective of the initial value of r or the initial state i_0. Here $\Pi T^{(\lambda)}(\Phi r^*) = \Phi r^*$. Further r^* satisfies*

$$\| \Phi r^* - J \|_D \leq \frac{1}{\sqrt{1-\beta_\lambda^2}} \| \Pi J - J \|_D$$

□

The initial value of the eligibility vector z_{-1} is irrelevant. Note that if J lies in the linear span of ϕ_k's, then $\tilde{J}(\cdot, r^*) = J$. In particular if $K = n$ then we have

$\tilde{J}(\cdot, r^*) = J$. In fact [54] proves the result only for aperiodic case. The proof for the general irreducible case follows from the results of the Chapter 3 and an analysis along the same lines as in [54]. We outline the proof below.

Let $X_t = (i_t, i_{t+1}, z_t, g_t)$. Then X_t is a Markov process which has a steady state distribution. From [54] the TD(λ) update may be written as

$$r_{t+1} = r_t + \gamma_t[A(X_t)r_t + b(X_t)]$$

where $A(X_t) = z_t(\beta\phi'(i_{t+1}) - \phi'(i_t))$ and $b(X_t) = z_t g_t$. Note that the probability transition kernel for the process X_t does not depend on r_t. If we denote by $\mathrm{E}_0[\cdot]$ the steady state expectation with respect to the invariant distribution of the Markov process X_t, then

$$\begin{aligned} A &= \mathrm{E}_0[A(X_t)] = \Phi' D(P^{(\lambda)} - I)\Phi \\ b &= \mathrm{E}_0[b(X_t)] = \Phi' D(I - \beta\lambda P)^{-1}\bar{g} \end{aligned}$$

A is negative definite and $Ar^* + b = \underline{0}$ (see [54]). Hence by Theorem 3.2 in Chapter 3 we have $r_t \to r^*$.

In Theorem 3.2 we may use the Lyapunov function $U(\theta) = \frac{1}{2}(\theta - \theta_*)'(\theta - \theta_*)$, where $\theta = r$ and $\theta_* = r^*$. All the assumptions of Theorem 3.2 are satisfied (see also Chapter 5).

Also the moment conditions on the immediate cost in Assumption 4.2(b) may be relaxed in that

$$\mathrm{E}\left[(|g_t|)^k \mid i_t = i\right] < \infty, \qquad \forall i \in \mathcal{S}$$

need be satisfied only upto a sufficiently large k ($k = 4$) and not for all $k > 0$ (see Chapter 3, Proposition 3.1).

4.5 TD(λ) For Learning

Here we are interested in learning the optimal value function and policy by reinforcement methods in an MDP framework. The notation for state space, action space, the transition probabilities and the assumptions on the immediate cost are as in Section 4.1. Neither the transition probabilities nor the distibution or expected value of immediate cost are known in advance. In this section we use i_t to denote the state at time t and a_t to denote the action taken at time t.

Assumption 4.3 *Let the following hold.*

(a). *The non-negative step sizes $\gamma_t \downarrow 0$ are pre-determined and satisfy*

$$\sum_{t=0}^{\infty} \gamma_t = \infty; \quad \sum_{t=0}^{\infty} \gamma_t^2 < \infty.$$

(b). *For each state action pair $(i,a) \in \mathcal{Q}$, let the pre-determined scalar non-negative step sizes $\gamma_t(i,a)$ be such that*

$$\sum_{t=1}^{\infty} \gamma_t(i,a) = \infty; \quad \sum_{t=1}^{\infty} (\gamma_t(i,a))^2 < \infty.$$

(c). *$K = n$ and*

$$\phi_k = e_k = [0, \cdots, 0, \overbrace{1}^{k^{\text{th}} \text{ entry}}, 0, \cdots, 0]^T$$

implying lookup table representation. Here e_k is the k^{th} standard basis vector in $\mathbf{R}^n$.

(d). *The immediate cost g_t has finite moments, i.e. $E[(|g_t|)^k \mid i_t = i, a_t = a] < \infty$; $\forall i \in \mathcal{S}, a \in \mathcal{A}(i), \forall k \in \mathbf{N}$.*

(e). *For some $\delta \in \Lambda_{\bar{\epsilon}}$ with $\underline{0} < \bar{\epsilon} \leq \tilde{\epsilon}$, assume that the Markov chain corresponding to the stationary policy δ is irreducible. In fact this implies that for any $\delta \in \Lambda_{\hat{\epsilon}}$, with $\underline{0} < \hat{\epsilon} \leq \tilde{\epsilon}$, the corresponding Markov chain is irreducible with the same period and has a unique invariant distribution π^{δ} with positive components, which depends on δ. Note that $\tilde{\epsilon}(i) = \frac{1}{|\mathcal{A}(i)|}$.*

□

Assumption 4.3(a) may be replaced by Assumption 4.2(a').

Assumption 4.3(e) is equivalent to the statement that any stationary fully randomized policy gives rise to an irreducible Markov chain, that is the MDP is communicating (see Section 5.2 in Chapter 5). Fix a policy $\delta \in \Lambda_{\bar{\epsilon}}, \ \ \underline{0} < \bar{\epsilon} \leq \tilde{\epsilon}$. We want to estimate J^{δ}, the cost to go for the stationary fully randomized policy δ and the Q values for policy δ given by $Q^{\delta}(i,a) = \mathsf{h}(i,a,J^{\delta}), \ \forall (i,a) \in \mathcal{Q}$. We use $\tilde{J}(\cdot, r) = \Phi r = r$ to approximate $J^{\delta}(\cdot)$. Note that Φ is the identity matrix. Let i_t and a_t be the state and action taken at time $t \in \mathbf{N}_0$, while using policy δ and let g_t be the corresponding immediate cost incurred. Note that our results handle any irreducible Markov chain, whether aperiodic or not.

Algorithm 4.1

$$\begin{aligned} z_{-1} &= \underline{0}, \\ \tau_{-1}(i,a) &= 0, \qquad i \in \mathcal{S}, a \in \mathcal{A}(i) \end{aligned}$$

The update rule is as follows (starting at $t = 0$)

$$z_t = (\beta\lambda) z_{t-1} + \underbrace{\phi(i_t)}_{e_{i_t}},$$

$$
\begin{aligned}
d_t &= g_t + \beta \tilde{J}(i_{t+1}, r_t) - \tilde{J}(i_t, r_t), \\
r_{t+1} &= r_t + \gamma_t d_t z_t, \\
\tau_t(i_t, a_t) &= \tau_{t-1}(i_t, a_t) + 1, \\
\tau_t(i, a) &= \tau_{t-1}(i, a), \qquad \forall (i, a) \neq (i_t, a_t), (i, a) \in \mathcal{Q} \\
Q_{t+1}(i_t, a_t) &= Q_t(i_t, a_t) + \gamma_{\tau_t(i_t, a_t)}(i_t, a_t) \cdot \\
&\quad \Big(g_t + \beta \tilde{J}(i_{t+1}, r_t) - Q_t(i_t, a_t) \Big) \\
Q_{t+1}(i, a) &= Q_t(i, a), \qquad \forall (i, a) \neq (i_t, a_t), (i, a) \in \mathcal{Q} \\
t &= t + 1
\end{aligned}
$$

□

$\tau_t(i, a)$ represents the number of times action a has been taken from state i by time $t \in \mathbf{N}_0$. Under Assumption 4.3, Lemma 4.3 ensures that $\tilde{J}(\cdot, r_t) \to J^\delta$ almost surely. In addition, since all state-action pairs in $\mathcal{Q}$ are "visited" infinitely often under policy δ, standard results from stochastic approximation theory [16] can be used to show that $Q_t \to Q^\delta$ almost surely. The convergence holds irrespective of the initialization of z, r and Q. All that is required of the non-negative step size parameters $\gamma_t(i, a)$ is that they should satisfy the standard assumptions

$$
\sum_{t=1}^{\infty} \gamma_t(i, a) = \infty; \quad \sum_{t=1}^{\infty} (\gamma_t(i, a))^2 < \infty
$$

almost surely and may be allowed to be random and can depend on the past history (at the time the step size is used).

Consider the following algorithm.

Algorithm 4.2 *Let* $\underline{0} < \bar{\epsilon}_k < \tilde{\epsilon}$ *be a sequence of positive vectors in* $\mathbf{R}^n$.

1. *Set* $k = 0$

2. *Select an arbitrary stationary deterministic policy* $\mu_0 \in \Upsilon$.

3. *Choose the stationary randomized extremal policy* $\delta_k \in \Lambda_{\bar{\epsilon}_k}$ *associated with* μ_k *and run Algorithm 4.1, for "large" random number of steps* n_k *till* Q_{n_k} *"nearly"converges to* Q^{δ_k} *and* $\tilde{J}(\cdot, r_{n_k})$ *"nearly" converges to* J^{δ_k}. *Set* $\tilde{Q} = Q_{n_k}$. *Let*

$$\varsigma_k = \max_{(i,a)\in\mathcal{Q}} |Q^{\delta_k}(i,a) - \tilde{Q}(i,a)|$$

4. *Set* $k = k + 1$ *and update the policy to* μ_k, *where*

$$\mu_k(i) = \arg\min_{a\in\mathcal{A}(i)} \tilde{Q}(i,a)$$

5. *Go to step* 3.

□

Theorem 4.1 *Consider the Algorithm 4.2 and let Assumption 4.3 hold. Then we have the following results*

1. *Given any scalar* $\epsilon > 0$, *there exists an* $\bar{\epsilon} \in \mathbf{R}^n$ *with* $\underline{0} < \bar{\epsilon} < \tilde{\epsilon}$ *and a number* $\varsigma > 0$ *such that if*

$$\limsup_{k\to\infty} \bar{\epsilon}_k(i) \;<\; \bar{\epsilon}(i), \qquad \forall i \in \mathcal{S}$$

and

$$\limsup_{k\to\infty} \varsigma_k \;<\; \varsigma,$$

then $\limsup_{k\to\infty} \| J^{\mu_k} - J^* \| < \epsilon$ *and* $\limsup_{k\to\infty} \| J^{\delta_k} - J^* \| < \epsilon$.

2. *Given any scalar* $\epsilon > 0$ *there exists* $\varsigma > 0$ *and* $\bar{\epsilon}$ *with* $\underline{0} < \bar{\epsilon} < \tilde{\epsilon}$ *such that if* $\bar{\epsilon}_k < \bar{\epsilon}$ *and* $\varsigma_k < \varsigma$, $\forall k$ *then* J^{μ_k} *converges to* J^* *in a finite number of steps* $(\leq |\Upsilon|)$ *and* $\| J^{\delta_k} - J^{\mu_k} \| < \epsilon \; \forall k$.

3. *In particular if* $\limsup_{k\to\infty} \bar{\epsilon}_k(i) = 0$, $\forall i \in \mathcal{S}$ *and* $\limsup_{k\to\infty} \varsigma_k = 0$ *then* $\| J^{\mu_k} - J^* \| \to 0$ *and* $\| J^{\delta_k} - J^* \| \to 0$. *In fact* $J^{\mu_k} = J^*$ *for all large* k.

□

Proof of Theorem 4.1

The fact that J^δ is a continuous function of δ on Λ implies that given any $\epsilon > 0$, there exists $\eta > 0$ (dependent on ϵ) such that $\| J^\mu - J^\delta \| < \epsilon$ for each $\mu \in \Upsilon$ and $\delta \in \Lambda$ with $\mathsf{d}(\mu, \delta) < \eta$. Also the extremal policies of $\Lambda_{\bar{\epsilon}}$, $\underline{0} < \bar{\epsilon} < \tilde{\epsilon}$ converges under the metric d to the corresponding deterministic policies as $\bar{\epsilon} \to \underline{0}$. These along with Corollary 4.2 and the comments following its statement proves the claims in Theorem 4.1.

□

The step size parameters used in step 3 of Algorithm 4.2 can vary for different policy evaluations (i.e. different ks). Our algorithm is somewhat similar in spirit to the Modified Q-learning in [42]. Finally the requirements on the existence of all moments of the immediate cost in Assumption 4.3(d) may be relaxed to the

requirement that $\mathrm{E}[(|g_t|)^k \mid i_t = i, a_t = a] < \infty,\ \forall (i,a) \in \mathcal{Q}$ need be satisfied only upto a sufficiently large k ($k = 4$) and not for all $k > 0$ (see Chapter 3, Proposition 3.1).

Note that instead of using an extremal policy $\delta_k \in \Lambda_{\bar{\epsilon}_k}$ to approximate μ_k, we could have chosen any $\tilde{\delta}_k \in \text{Interior}(\Lambda)$ such that $[\tilde{\delta}_k(i)]_{\mu_k(i)} \geq (1 - (|\mathcal{A}(i)| - 1)\bar{\epsilon}_k(i)),\ \forall i \in \mathcal{S}$; for instance $\tilde{\delta}_k$ could be made to depend on the approximation to $Q^{\tilde{\delta}_{k-1}}$ obtained in the previous step. We note that the initial condition i_0, r_0, z_{-1} and Q_0 when calling Algorithm 4.1 in step 3 of Algorithm 4.2 may be arbitray, but can be set to the final values obtained in the previous iteration.

In [16], a particular learning scheme uses TD(λ) to approximate the value functions for deterministic stationary policies, before trying to estimate the corresponding Q-values by further simulation, in order to obtain a policy update. This may lead to problems with convergence when we are using online schemes with arbitrary initialization. This methodology thus differs from the one proposed in this paper, where we deal with stationary randomized policies, and on-line updates of the Q-values, along with the TD updates.

Now we give an example which shows the problem associated with arbitrary initialization when we use only deterministic policies. Consider the following two state MDP where $\mathcal{S} = \{1, 2\},\ \mathcal{A}(1) = \{1, 2\},\ \mathcal{A}(2) = \{1\}$. Let the discount factor be β, where $0 < \beta < 1$. At state 1, there are two options: under the first, say $u = 1$, we stay at state 1 with probability 1 and a cost $M > 0$ is incurred; under the second, say $u = 2$, we move to state 2 with probability 1 at zero cost. From state 2, there is only one possible action, say $u = 1$, under which we move to state 1 with

probability 1 at zero cost.

The optimal action at state 1 is to use action 2 . The corresponding optimal cost is $J^*(1) = J^*(2) = 0$. The only other possible stationary deterministic policy is $\mu = (1 \;\; 1)'$, the one corresponding to using action 1 in state 1, and the corresponding cost to go is $J^\mu(1) = \frac{M}{(1-\beta)}$ and $J^\mu(2) = \beta\frac{M}{(1-\beta)}$. Suppose that we initialize the algorithm with $J_0(1) = \alpha_1$ and $J_0(2) = \alpha_2$ with $\alpha_1 \leq \frac{M}{(1-\beta)}$ and $\alpha_2 > \frac{1}{\beta}\frac{M}{(1-\beta)} + \alpha$, α a large positive number.

The corresponding greedy policy is $\mu = (1 \;\; 1)'$. Let the initial state be $i_0 = 1$. This means that the system stays at state 1 and $J(2)$ is not updated, whereas $J(1)$ converges to $\frac{M}{(1-\beta)}$. The greedy policy for this remains the same $\mu = (1 \;\; 1)'$. Thus, as long as we start in state 1, the greedy policy does not visit state 2, and the value of $J(2)$ never changes, and we are stuck with a non-optimal policy.

Now we give an example of an MDP where under any stationary policy (deterministic or randomized) the corresponding Markov chain is periodic (not aperiodic). We consider a finite one dimensional random walk with n states ($n \geq 3$); $\mathcal{S} = \{1, 2, \ldots, n\}$. The feasible actions are given by $\mathcal{A}(1) = \{1\} = \mathcal{A}(n),\ \mathcal{A}(i) = \{1, 2\}$ for $1 < i < n$. The transition probabilities for states $1 < i < n$ are given by $p_{i,i+1}(1) = 1 - p,\ p_{i,i-1}(1) = p$, where $1 > p > \frac{1}{2}$ and $p_{i,i+1}(2) = q,\ p_{i,i-1}(2) = 1 - q$, where $1 > q > \frac{1}{2}$. Also $p_{1,2}(1) = 1,\ p_{n,n-1}(1) = 1$. The immediate costs are given by $g(i, a, j) = c(j)$ for $i, j \in \mathcal{S}$ and $a \in \mathcal{A}(i)$. Also

$$\begin{aligned} c(j) &= c(n - j + 1) \qquad \text{for } j = 1, \ldots, \left\lfloor \frac{n}{2} \right\rfloor, \\ c(j+1) &< c(j) \qquad\qquad\quad \text{for } j = 1, \ldots, \left\lceil \frac{n}{2} \right\rceil - 1. \end{aligned}$$

Here under any stationary policy (deterministic or randomized) the Markov chain is irreducible and has period 2. $c(i)$ might represent the temperature of state $i \in \mathcal{S}$. We are interested in moving toward states with lower temperature.

Chapter 5

Temporal Difference Schemes For Average Cost MDPs

In this chapter we propose a reinforcement learning scheme for finding optimal and sub-optimal policies for the finite state, finite action, Markov Decision Problem (MDP) with the average cost criterion [4, 12, 40]. Online learning is utilized along with temporal difference schemes for approximating differential cost functions to obtain a direct adaptive control scheme for the MDP. We provide convergence results of the algorithm for unichain MDP with a common recurrent state [12, 40]. In particular we do not assume that the recurrent class is aperiodic as in [55], for any stationary policy.

In Section 5.1 we revisit the average cost MDP model. Section 5.2 deals with a classification scheme of MDPs. In Section 5.3 we discuss some properties of the transition probability matrix and its application to Markov Cost process. In Section 5.4 we deal with Bellman equation and policy iteration schemes for unichain MDP with a common recurrent state. Continuity issues of the limiting and differential matrices are dealt with in Section 5.5. In Section 5.6 we deal with approximate policy iteration for average cost MDPs. Section 5.7 deals with temporal difference schemes for estimating the average cost and differential cost (with linear function approximation) of a Markov Cost process. In particular we extend the results in[55] to Markov chains which are not necessarily aperiodic. Corresponding variations in

the proofs of the sub-results leading to the main results are dealt with. In particular we use the general stochastic approximation algorithm in Chapter 3. TD schemes are used in conjunction with on-line estimates of Q-values to solve the average cost MDP in Section 5.9

5.1 Average Cost MDP Model Revisited

We refer the reader to Section 4.1 and Section 4.2 of Chapter 4 for notations regarding the homogeneous MDP model, the admissible policies and stationary randomized policies. However in this chapter we are dealing with the average cost problem. The changes in the notation are dealt with in the appropriate sections of the chapter. In particular we assume the state space to be $\mathcal{S} = \{1, 2, \ldots, n\}$, where n is a positive integer. The state of the system at time $t \in \mathbf{N}_0$ is denoted by s_t, which is an element of $\mathcal{S}$. Here $\mathbf{N}_0$ denotes the set of non-negative integers. The action taken at time t, is denoted by u_t, where $u_t \in \mathcal{A}(s_t)$. $\mathcal{A}(i) = \{1, 2, \ldots, |\mathcal{A}(i)|\}$ denotes the nonempty finite control constraint set for $i \in \mathcal{S}$. Let $p_{ij}(a)$ denote the transition probability from state i to state j, when action a is taken from state i at any time $t \in \mathbf{N}_0$, for $i \in \mathcal{S}$ and $a \in \mathcal{A}(i)$. The immediate cost incurred at time $t \in \mathbf{N}_0$, while taking action u_t from state s_t is denoted by g_t, where $u_t \in \mathcal{A}(s_t)$. The probability distribution of g_t might depend on $s_t,\ s_{t+1}$ and u_t, but not on the past history up to time t, given $s_t,\ s_{t+1}$ and u_t. Let $g(i, a) = E[g_t | s_t = i, u_t = a]$ denote the expected value of the immediate cost for taking action a from state i at any time t, for $i \in \mathcal{S}$, $a \in \mathcal{A}(i)$. We assume that the expected immediate costs

have finite (hence bounded) second moments; i.e. $\mathrm{E}\left[(|g_t|)^2 \mid s_t = i, u_t = a\right] < \infty$ for $i \in \mathcal{S},\ a \in \mathcal{A}(i)$. The set of all admissible policies is denoted by $\mathcal{M}$. The set of all stationary deterministic policies is denoted by Υ, and the set of all stationary randomized policies is denoted by Λ. We use the terminology, (fully) randomized stationary policies and stationary (fully) randomized policies interchangeably.

The performance criterion for the average cost problem is the well defined quantity $\bar{\vartheta}^{\nu} \in \mathbf{R}^n$, given by

$$\bar{\vartheta}^{\nu}(i) \;=\; \limsup_{N\to\infty} \frac{1}{N}\mathrm{E}^{\nu}\left[\sum_{t=0}^{N-1} g_t \mid s_0 = i\right], \tag{5.1}$$

when the policy $\nu \in \mathcal{M}$ is used and the initial state is $i \in \mathcal{S}$. The aim is to find a policy $\nu^* \in \mathcal{M}$ such that $\bar{\vartheta}^{\nu^*}(i) = \bar{\vartheta}^*(i),\ \forall i \in \mathcal{S}$. Here $\bar{\vartheta}^* \in \mathbf{R}^n$, given by

$$\bar{\vartheta}^*(i) = \inf_{\nu\in\mathcal{M}} \bar{\vartheta}^{\nu}(i), \qquad i \in \mathcal{S}$$

is the optimal average cost function or vector. It is well known that there exists a stationary deterministic policy [12, 40] which is optimal. We would like to add that (see [40]),

$$\inf_{\nu\in\mathcal{M}} \limsup_{N\to\infty} \frac{1}{N}\mathrm{E}^{\nu}\left[\sum_{t=0}^{N-1} g_t \mid s_0 = i\right] = \inf_{\nu\in\mathcal{M}} \liminf_{N\to\infty} \frac{1}{N}\mathrm{E}^{\nu}\left[\sum_{t=0}^{N-1} g_t \mid s_0 = i\right], \qquad i \in \mathcal{S}$$

Also note that the limit exists in equation 5.1, for any stationary policy $\delta \in \Lambda$, i.e.

$$\bar{\vartheta}^{\delta}(i) \;=\; \lim_{N\to\infty} \frac{1}{N}\mathrm{E}^{\delta}\left[\sum_{t=0}^{N-1} g_t \mid s_0 = i\right], \qquad \forall i \in \mathcal{S}.$$

5.2 Classification Of MDPs

We may classify MDPs as follows [40].

Any MDP is referred to as *general*.

An MDP is called *recurrent* or *ergodic*, if the transition probability matrix corresponding to every stationary deterministic policy gives rise to an irreducible Markov chain, i.e. has only one recurrent class which encompasses all the states.

An MDP is called *unichain*, if the transition probability matrix corresponding to every stationary deterministic policy is unichain, that is, the corresponding Markov chain consists of a single recurrent class plus a possibly empty set of transient states.

An MDP is called *communicating*, if for every pair of states i and j in $\mathcal{S}$, there exists a stationary deterministic policy $\mu \in \Upsilon$ (depending on i and j) under which j is accessible from i, that is $[P_\mu^k]_{ij} > 0$ for some $k \geq 1$. (In fact if such a k exists, there exists an l, with $1 \leq l \leq n$, such that $[P_\mu^l]_{ij} > 0$). Here P_μ is the transition probability matrix associated with stationary deterministic policy μ. In other words an MDP is a communicating MDP, if for any stationary fully randomized policy $\delta \in \Lambda$, the corresponding Markov chain is irreducible (see Appendix B).

An MDP is called *weakly communicating*, if there exists a closed set of states, with each state in that set accessible from every other state in the set, under some stationary deterministic policy, plus a possibly empty set of states which is transient under every policy. In other words an MDP is weakly communicating, if for any stationary fully randomized policy δ, the corresponding Markov chain has a single recurrent class (i.e. δ is a unichain policy) and every transient state (if it exists) of δ are transient under every policy. Note that it may be shown that for a set of states to be transient under every policy, it is sufficient that they be transient under every

stationary deterministic policy.

An MDP is called *multichain*, if the Markov chain corresponding to at least one stationary deterministic policy contains two or more closed irreducible recurrent classes.

Note that a recurrent MDP is unichain as well as communicating. A communicating MDP is weakly communicating and a unichain MDP is weakly communicating. In fact (see Appendix B) for any stationary fully randomized policy of a unichain MDP, the unique recurrent class is the union of recurrent classes of all the stationary deterministic policies.

5.3 Some Properties Of The Transition Probability Matrix

5.3.1 Basics

Here we give an important result stated in [12, Proposition 4.1.1].

Lemma 5.1 *For any stochastic matrix P and $\beta \in [0,1)$ there holds*

$$(I - \beta P)^{-1} = (1-\beta)^{-1} P^* + L + \mathcal{O}(|1-\beta|),$$

where $\mathcal{O}(|1-\beta|)$ is a β-dependent matrix such that

$$\lim_{\beta \to 1} \mathcal{O}(|1-\beta|) = \mathbf{0},$$

and the matrix P^* and L are given by

$$P^* = \lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} P^k, \tag{5.2}$$

$$L = (I - P + P^*)^{-1} - P^*. \tag{5.3}$$

The limit and the inverse, in the above equations exist. Furthermore P^* and L satisfy the following equations:

$$\begin{aligned} P^* &= PP^* = P^*P = P^*P^*, \\ P^*L &= LP^* = \mathbf{0}, \qquad (5.4) \\ P^* + L &= I + PL. \qquad (5.5) \end{aligned}$$

□

Here $\mathbf{0}$ is the zero matrix. Note that

$$(I - P + P^*)(I - P^*) = (I - P + P^*) - P^*,$$

and hence

$$(I - P^*) = I - (I - P + P^*)^{-1}P^*,$$

which implies

$$(I - P + P^*)^{-1}P^* = P^*.$$

Hence

$$\begin{aligned} L &= (I - P + P^*)^{-1} - P^* \\ &= (I - P + P^*)^{-1}(I - P^*). \end{aligned}$$

Let $\{M_l : l \geq 0\}$ be a sequence of $n \times n$ real valued matrices. We say that M is a Cesaro limit of $\{M_l : l \geq 0\}$ if

$$\lim_{N\to\infty} \frac{1}{N} \sum_{l=0}^{N-1} M_l = M,$$

and we write

$$C - \lim_{N\to\infty} M_N = M.$$

Note that if the limit of the sequence $\{M_l : l \geq 0\}$ exists, that limit is the Cesaro limit. See [40].

Note that P^* is called the *limiting matrix* corresponding to P, with [40]

$$P^* = C - \lim_{N \to \infty} P^N.$$

L is called the *differential matrix* corresponding to P, also known as the *Drazin Inverse* of $(I - P)$. See [40].

Lemma 5.2 *We have*

$$L = C - \lim_{N \to \infty} \sum_{k=0}^{N-1} (P^k - P^*),$$

where $P^0 = I$.

□

Proof of Lemma 5.2

To prove this we need to prove

$$(I - P + P^*)^{-1} = C - \lim_{N \to \infty} \sum_{k=0}^{N-1} (P - P^*)^k,$$

with $(P - P^*)^0 = I$. Note that $(P - P^*)^k = P^k - P^*$ for $k \geq 1$.

Note that in the following,

$$\sum_{k=i}^{j} f_k = 0,$$

if $j < i$, by convention, where f_k is some real valued quantity indexed by integer k. Now

$$(I - P + P^*) \left(\frac{1}{N} \sum_{l=0}^{N-1} \sum_{k=0}^{l} (P - P^*)^k \right)$$

$$
\begin{aligned}
&= \frac{1}{N}\sum_{l=0}^{N-1}\left((I-P+P^*)+\sum_{k=1}^{l}(P^k-P^{k+1}+P^*-P^*)\right)\\
&= \frac{1}{N}\left[N(I-P+P^*)+\sum_{l=1}^{N-1}\sum_{k=1}^{l}(P^k-P^{k+1})\right]\\
&= \frac{1}{N}\left[N(I-P+P^*)+\sum_{l=1}^{N-1}(P-P^{l+1})\right]\\
&= I-P+P^*+\frac{(N-1)}{N}P-\frac{(N-1)}{N}\frac{1}{(N-1)}\sum_{l=1}^{N-1}P^{l+1}
\end{aligned}
$$

Note that

$$
\lim_{N\to\infty}\left(I-P+P^*+\frac{(N-1)}{N}P-\frac{(N-1)}{N}\frac{1}{(N-1)}\sum_{l=1}^{N-1}P^{l+1}\right)=I.
$$

Since $(I-P+P^*)$ is invertible,

$$
\begin{aligned}
C-\lim_{N\to\infty}\sum_{k=0}^{N-1}(P-P^*)^k &= (I-P+P^*)^{-1}I\\
&= (I-P+P^*)^{-1}.
\end{aligned}
$$

□

5.3.2 Application To Markov Cost Process

Define the expected immediate cost vector $\bar{g}^\delta$ and transition probability matrix P_δ, corresponding to stationary randomized policy $\delta\in\Lambda$, by

$$
\begin{aligned}
\bar{g}^\delta(i) &= \sum_{a\in\mathcal{A}(i)}[\delta(i)]_a\, g(i,a),\\
[P_\delta]_{ij} \equiv p_{ij}^\delta &= \sum_{a\in\mathcal{A}(i)}[\delta(i)]_a\, p_{ij}(a).
\end{aligned}
$$

Let $J_{\beta,\delta}$ denote the infinite horizon discounted cost corresponding to stationary randomized policy $\delta\in\Lambda$ and discount factor $\beta\in[0,1)$. That is

$$
J_{\beta,\delta} = \sum_{k=0}^{\infty}\beta^k P_\delta^k\bar{g}^\delta
$$

$$= \left(\sum_{k=0}^{\infty} \beta^k P_\delta^k\right) \bar{g}^\delta = (I - \beta P_\delta)^{-1} \bar{g}^\delta.$$

The following proposition follows from Lemma 5.1 and relates the β discounted cost and average cost corresponding to a stationary policy (see [12, Proposition 4.1.2, page 182]).

Proposition 5.1 (Truncated Laurent Series Expansion) *For any stationary policy $\delta \in \Lambda$ and $\beta \in [0, 1)$*

$$J_{\beta,\delta} = (1-\beta)^{-1} \bar{\vartheta}^\delta + J_\delta^* + \mathcal{O}(|1-\beta|), \tag{5.6}$$

where $\bar{\vartheta}^\delta$ and J_δ^* are given by

$$\bar{\vartheta}^\delta = P_\delta^* \bar{g}^\delta,$$

and

$$J_\delta^* = L_\delta \bar{g}^\delta,$$

with

$$P_\delta^* = \left(\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} P_\delta^k\right),$$
$$L_\delta = (I - P_\delta + P_\delta^*)^{-1} - P_\delta^*.$$

Furthermore

$$\bar{\vartheta}^\delta = P_\delta \bar{\vartheta}^\delta,$$
$$\bar{\vartheta}^\delta + J_\delta^* = \bar{g}^\delta + P_\delta J_\delta^*.$$

□

Here $\mathcal{O}(|1-\beta|)$ is a β dependent vector, such that $\lim_{\beta\to 1}\mathcal{O}(|1-\beta|) = \underline{0}$, the zero vector.

Equation 5.6 is referred to as the *Truncated Laurent Series Expansion* of the discounted cost of a stationary policy δ. The vectors J^*_δ and $\bar{\vartheta}^\delta$ in the Truncated Laurent series expansion are uniquely defined, and will be referred to as the *bias* and *gain* of δ, respectively.

If P is the transition probability matrix corresponding to a unichain Markov chain, then

$$P^* = \underline{1}\pi',$$

where $\pi \in \mathbf{R}^n$ is the unique invariant distribution corresponding to P, and $\underline{1} \in \mathbf{R}^n$ is the vector with all components equal to one. Note that $\pi(i) = 0$, if i is a transient state and $\pi(i) > 0$, if i is an element of the unique recurrent class. Also $\sum_{i=1}^n \pi(i) = 1$. Note that $\pi' P = \pi'$. Also note that the eigen value 1 has an algebraic multiplicity of one and that any left eigen vector of P corresponding to eigen value 1 is a scalar multiple of π.

Let the operators $\bar{T}_\delta : \mathbf{R}^n \to \mathbf{R}^n$ and $\bar{T} : \mathbf{R}^n \to \mathbf{R}^n$ be defined as

$$(\bar{T}_\delta J)(i) \;=\; \bar{g}^\delta(i) + \sum_{j=1}^n p^\delta_{ij} J(j), \qquad \text{for } i = 1, 2, \ldots, n.$$

and

$$(\bar{T} J)(i) \;=\; \min_{a\in\mathcal{A}(i)} \left[g(i,a) + \sum_{j=1}^n p_{ij}(a) J(j) \right], \qquad \text{for } i = 1, 2, \ldots, n.$$

Here $J \in \mathbf{R}^n$. In particular for a stationary deterministic policy μ ($\in \Upsilon$), we have,

$$(\bar{T}_\mu J)(i) \;=\; g(i,\mu(i)) + \sum_{j=1}^n p_{ij}(\mu(i)) J(j), \qquad \text{for } i = 1, 2, \ldots, n.$$

We have the following lemma. See [12, Proposition 4.2.4] and [40].

Lemma 5.3 *Let $\delta \in \Lambda$ be a stationary policy which is unichain.*

(a) *Then there exists a scalar ϑ^δ and a vector $J_\delta \in \mathbf{R}^n$, such that*

$$\vartheta^\delta \underline{1} + J_\delta = \bar{T}_\delta J_\delta.$$

(b) *Let k be a fixed state. Then the system of equations*

$$\vartheta \underline{1} + J = \bar{T}_\delta J \qquad \textit{and} \quad J(k) = 0,$$

in the $(n+1)$ unknowns $\vartheta, J(1), J(2), \ldots, J(n)$; has a unique solution, with $\vartheta = \vartheta^\delta$. □

Note that the average cost starting at any state $i \in \mathcal{S}$ under policy δ is $\bar{\vartheta}^\delta(i) = \vartheta^\delta$. In the above lemma $\vartheta^\delta = (\pi^\delta)'\bar{g}^\delta$, where π^δ is the unique invariant distribution for the policy δ, and $\bar{g}^\delta$ is the expected immediate cost vector for policy δ. Note that if $L_\delta = (I - P_\delta + P_\delta^*)^{-1} - P_\delta^*$, where

$$P_\delta^* = \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} P_\delta^k,$$

then from equation 5.5,

$$P_\delta^* + L_\delta = I + P_\delta L_\delta,$$

and hence

$$P_\delta^* \bar{g}^\delta + L_\delta \bar{g}^\delta = \bar{g}^\delta + P_\delta L_\delta \bar{g}^\delta.$$

Now $P_\delta^* = \underline{1}(\pi^\delta)'$. With $\vartheta^\delta = (\pi^\delta)'\bar{g}^\delta$ and $J_\delta^* = L_\delta \bar{g}^\delta$, the above equation can be written as

$$\vartheta^\delta \underline{1} + J_\delta^* = \bar{T}_\delta J_\delta^*.$$

Note that $(\pi^\delta)' J^*_\delta = (\pi^\delta)' L_\delta \bar{g}^\delta = 0$, since $P^*_\delta L_\delta = \mathbf{0}$, from equation 5.4.

ϑ^δ is called the average cost (gain) and $L_\delta \bar{g}^\delta$ is called the *basic* differential cost (bias) for the Markov Cost process corresponding to policy δ. Note that a scalar ϑ and a vector J satisfies

$$\vartheta \underline{1} + J = \bar{T}_\delta J,$$

if and only if $\vartheta = \vartheta^\delta$ and $sp(J - L_\delta \bar{g}^\delta) = 0$, where, for a vector J the span semi-norm [40],

$$sp(J) \equiv \max_{i \in \mathcal{S}} J(i) - \min_{i \in \mathcal{S}} J(i).$$

Notice that as stated in Chapter 1, if $\vartheta \in \mathbf{R}$ and $J \in \mathbf{R}^n$ satisfy

$$\vartheta \underline{1} + J = \bar{T} J, \tag{5.7}$$

then $\vartheta = \vartheta^*$, the optimal average cost starting from any state $i \in \mathcal{S}$. In addition if $\delta \in \Lambda$ is any stationary policy such that $\bar{T}_\delta J = \bar{T} J$, then δ is average cost optimal.

Note that for vector $J, \bar{J} \in \mathbf{R}^n$, $sp(J - \bar{J}) = 0$ if and only if $J = \bar{J} + \alpha \underline{1}$ for some $\alpha \in \mathbf{R}$.

5.4 Unichain MDP With A Common Recurrent State

5.4.1 Bellman Equation

Assumption 5.1 *The MDP is unichain and one of the states, say 's', is such that it is recurrent under all stationary deterministic policies.*

□

We have the following variant of [11, Proposition 7.4.1].

Lemma 5.4 *Under Assumption 5.1, the following hold for the average cost per stage problem.*

(a) *The optimal average cost is the same for all initial states and together with some vector* $J^\diamond = (J^\diamond(1), J^\diamond(2), \ldots, J^\diamond(n))'$ *satisfies the Bellman Equation:*

$$\vartheta^* \underline{1} + J^\diamond = \bar{T} J^\diamond, \tag{5.8}$$

that is

$$\vartheta^* + J^\diamond(i) = \min_{a \in \mathcal{A}(i)} \left[g(i,a) + \sum_{j=1}^{n} p_{ij}(a) J^\diamond(j) \right], \qquad \text{for } i = 1, 2, \ldots, n.$$

Furthermore, if $\mu(i)$ *attains the minimum in the above equation for all* i*, the stationary policy* μ *is optimal. Fix* $k \in \{1, 2, \ldots, n\}$*. Then in addition, out of all vectors* $J^\diamond$ *satisfying this equation, there is a unique vector for which* $J^\diamond(k) = 0$*.*

(b) *A scalar* ϑ *and a vector* $J = (J(1), J(2), \ldots, J(n))'$ *satisfy Bellman's equation if and only if* $\vartheta = \vartheta^*$ *(the optimal average cost for each initial state) and* $sp(J^\diamond - J) = 0$*.*

(c) *Fix* $k \in \{1, 2, \ldots, n\}$*. Given a stationary deterministic policy* μ*, with corresponding average cost per stage* ϑ^μ*, there is a unique vector* $J_\mu \in \mathbf{R}^n$*, such that* $J_\mu(k) = 0$ *and*

$$\vartheta^\mu + J_\mu(i) = g(i, \mu(i)) + \sum_{j=1}^{n} p_{ij}(\mu(i)) J_\mu(j), \qquad i = 1, 2, \ldots, n.$$

□

We may characterize $J^\diamond$ in part (a) of Lemma 5.4 as follows. Let μ^* be any Blackwell optimal policy [12, 40]. Then by [12, Propostion 4.1.4] (see it's proof) the scalar $\vartheta^* = (\pi^{\mu^*})'\bar{g}^{\mu^*}$ and vector $L_{\mu^*}\bar{g}^{\mu^*}$ satisfy the Bellman equation. All the Blackwell optimal policies have the same gain and bias.

Note that $\vartheta \in \mathbf{R}$ and $J \in \mathbf{R}^n$ satisfy Bellman's equation if and only if $\vartheta = \vartheta^*$ and $sp(J - L_{\mu^*}\bar{g}^{\mu^*}) = 0$. Also note that if there is a unique policy μ which minimizes the RHS (Right Hand Side) of the Bellman equation, then it is a Blackwell Optimal policy (unique Blackwell optimal policy).

5.4.2 Policy Iteration

Let μ and $\bar{\mu}$ be unichain policies, and ϑ^μ and $\vartheta^{\bar{\mu}}$ be the average cost (independent of initial state) corresponding to policies μ and $\bar{\mu}$. Let J_μ and $J_{\bar{\mu}}$ be differential cost vectors for μ and $\bar{\mu}$ which satisfy

$$\vartheta^\mu \underline{1} + J_\mu = \bar{g}^\mu + P_\mu J_\mu = \bar{T}_\mu J_\mu, \tag{5.9}$$

and

$$\vartheta^{\bar{\mu}} \underline{1} + J_{\bar{\mu}} = \bar{g}^{\bar{\mu}} + P_{\bar{\mu}} J_{\bar{\mu}} = \bar{T}_{\bar{\mu}} J_{\bar{\mu}}. \tag{5.10}$$

Let

$$\begin{aligned} P^*_\mu &= \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} P^k_\mu, \\ P^*_{\bar{\mu}} &= \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} P^k_{\bar{\mu}}. \end{aligned}$$

Lemma 5.5 *Let μ and $\bar{\mu}$ be stationary deterministic policies that are unichain.*

Then

$$(\vartheta^{\mu} - \vartheta^{\bar{\mu}})\underline{1} = P^{*}_{\bar{\mu}} \underbrace{(\vartheta^{\mu}\underline{1} + J_{\mu} - \bar{g}^{\bar{\mu}} - P_{\bar{\mu}}J_{\mu})}_{\bar{T}_{\mu}J_{\mu} - \bar{T}_{\bar{\mu}}J_{\mu}}$$

□

Proof of Lemma 5.5

Note that from equation 5.9 and equation 5.10

$$(\vartheta^{\mu}\underline{1} + J_{\mu} - \bar{g}^{\bar{\mu}} - P_{\bar{\mu}}J_{\mu}) \;=\; (\vartheta^{\mu} - \vartheta^{\bar{\mu}})\underline{1} + (J_{\mu} - J_{\bar{\mu}}) - P_{\bar{\mu}}(J_{\mu} - J_{\bar{\mu}}).$$

Multiplying the above relation with $P^{k}_{\bar{\mu}}$ and adding from $k = 0$ to $N-1$, we obtain,

$$\sum_{k=0}^{N-1} P^{k}_{\bar{\mu}}(\vartheta^{\mu}\underline{1} + J_{\mu} - \bar{g}^{\bar{\mu}} - P_{\bar{\mu}}J_{\mu}) \;=\; N(\vartheta^{\mu} - \vartheta^{\bar{\mu}})\underline{1} + (J_{\mu} - J_{\bar{\mu}}) - P^{N}_{\bar{\mu}}(J_{\mu} - J_{\bar{\mu}}).$$

Divide by N and taking the limit as $N \to \infty$ we obtain

$$P^{*}_{\bar{\mu}} \overbrace{(\vartheta^{\mu}\underline{1} + J_{\mu} - \bar{g}^{\bar{\mu}} - P_{\bar{\mu}}J_{\mu})}^{(\bar{T}_{\mu}J_{\mu} - \bar{T}_{\bar{\mu}}J_{\mu})} \;=\; (\vartheta^{\mu} - \vartheta^{\bar{\mu}})\underline{1}.$$

□

Corollary 5.1 *Let μ and $\bar{\mu}$ be unichain stationary deterministic policies. If $\bar{T}_{\bar{\mu}}J_{\mu} = \bar{T}J_{\mu}$, then $\vartheta^{\bar{\mu}} \leq \vartheta^{\mu}$.*

□

This follows from the fact that

$$\bar{T}_{\bar{\mu}}J_{\mu} = \bar{T}J_{\mu} \leq \bar{T}_{\mu}J_{\mu}.$$

Notice that for any $\alpha \in \mathbf{R}$ and any stationary policy $\delta \in \Lambda$ and any $J \in \mathbf{R}^{n}$,

$$\bar{T}_{\delta}(J + \alpha\underline{1}) \;=\; \bar{T}_{\delta}J + \alpha\underline{1},$$

and

$$\bar{T}(J + \alpha\underline{1}) \;=\; \bar{T}J + \alpha\underline{1}.$$

Lemma 5.6 *Let μ and $\bar{\mu}$ be unichain stationary deterministic policies. Let ϑ^{μ} and $\vartheta^{\bar{\mu}}$ be the corresponding average costs, J_{μ} and $J_{\bar{\mu}}$ be corresponding differential costs satisfying*

$$\vartheta^{\mu}\underline{1} + J_{\mu} \;=\; \bar{T}_{\mu}J_{\mu},$$

and

$$\vartheta^{\bar{\mu}}\underline{1} + J_{\bar{\mu}} \;=\; \bar{T}_{\bar{\mu}}J_{\bar{\mu}}$$

Also let $\bar{T}J_{\mu} = \bar{T}_{\bar{\mu}}J_{\mu}$ and $sp(J_{\mu} - J_{\bar{\mu}}) = 0$. Then

$$\bar{T}_{\bar{\mu}}J_{\bar{\mu}} = \bar{T}J_{\bar{\mu}}$$

and hence $\vartheta^{\bar{\mu}} = \vartheta^{*}$, i.e. $\bar{\mu}$ is an optimal policy.

□

Proof of Lemma 5.6

Now $J_{\bar{\mu}} - J_{\mu} = \alpha\underline{1}$ for some scalar α since $sp(J_{\mu} - J_{\bar{\mu}}) = 0$.

$$\begin{aligned}
\vartheta^{\bar{\mu}} + J_{\bar{\mu}} &= \bar{T}_{\bar{\mu}}J_{\bar{\mu}} \\
&= \bar{T}_{\bar{\mu}}(J_{\mu} + \alpha\underline{1}) = \bar{T}(J_{\mu} + \alpha\underline{1}) \\
&= \bar{T}J_{\bar{\mu}}.
\end{aligned}$$

implying that $\vartheta^{\bar{\mu}} = \vartheta^{*}$. See Chapter 1 and [12, Proposition 4.2.1 and Proposition 4.2.2]. □

Lemma 5.7 *Let μ be a unichain policy with average cost $\vartheta^\mu \in \mathbf{R}$ and differential cost J_μ satisfying*

$$\vartheta^\mu \underline{1} + J_\mu = \bar{T}_\mu J_\mu = \bar{T} J_\mu$$

Let $\bar{\mu}$ be any unichain policy (with average cost $\vartheta^{\bar{\mu}}$ and differential cost $J_{\bar{\mu}}$ satisfying $\vartheta^{\bar{\mu}} \underline{1} + J_{\bar{\mu}} = \bar{T}_{\bar{\mu}} J_{\bar{\mu}}$) such that

$$\bar{T}_{\bar{\mu}} J_\mu = \bar{T} J_\mu$$

Then $\vartheta^{\bar{\mu}} = \vartheta^\mu = \vartheta^*$ and $sp(J_{\bar{\mu}} - J_\mu) = 0$. Also $\bar{T}_{\bar{\mu}} J_{\bar{\mu}} = \bar{T} J_{\bar{\mu}}$. □

Proof of Lemma 5.7

$\vartheta^\mu = \vartheta^*$ follows immediately from

$$\vartheta^\mu \underline{1} + J_\mu = \bar{T} J_\mu = \bar{T}_\mu J_\mu.$$

Now since $\bar{T}_{\bar{\mu}} J_\mu = \bar{T} J_\mu$ we have

$$\vartheta^\mu \underline{1} + J_\mu = \bar{T}_\mu J_\mu = \bar{T} J_\mu = \bar{T}_{\bar{\mu}} J_\mu.$$

Hence by Lemma 5.3, we have $\vartheta^\mu = \vartheta^{\bar{\mu}}$ and $sp(J_\mu - J_{\bar{\mu}}) = 0$. Since $J_{\bar{\mu}} = \alpha \underline{1} + J_\mu$, for some scalar α, we have

$$\begin{aligned} \bar{T}_{\bar{\mu}} J_{\bar{\mu}} &= \bar{T}_{\bar{\mu}} (J_\mu + \alpha \underline{1}) \\ &= \bar{T}_{\bar{\mu}} J_\mu + \alpha \underline{1} = \bar{T} J_\mu + \alpha \underline{1} \\ &= \bar{T} (J_\mu + \alpha \underline{1}) = \bar{T} J_{\bar{\mu}}. \end{aligned}$$

□

Let Assumption 5.1 hold. We now give the *policy iteration algorithm* [11, pages 432–435]. It operates as follows.

Given a stationary deterministic policy, we obtain an improved stationary deterministic policy by means of a minimization process, until no further improvement is possible. In particular at the typical step of the algorithm, we have a stationary deterministic policy μ_k. We then perform a policy evaluation step; that is, we obtain corresponding average and differential costs $\vartheta_k \in \mathbf{R}$ and $J_k \in \mathbf{R}^n$ satisfying

$$\vartheta_k \underline{1} + J_k = \bar{T}_{\mu_k} J_k \qquad \text{with } J_k(s) = 0.$$

Here 's' is the common recurrent state. We subsequently perform a policy improvement step; that is, we find stationary deterministic policy μ_{k+1} such that

$$\bar{T}_{\mu_{k+1}} J_k = \bar{T} J_k$$

If $\vartheta_{k+1} = \vartheta_k$ and $J_{k+1} = J_k$ then the algorithm terminates; otherwise the process is repeated with μ_{k+1} replacing μ_k. To prove that the policy iteration algorithm terminates, it is sufficient that each iteration makes some irreversible progress towards optimality, since there are finitely many stationary deterministic policies. The following proposition [11, Proposition 7.4.2] shows the type of irreversible progress we can demonstrate. It also shows that an optimal policy is obtained upon termination.

Proposition 5.2 (Policy Iteration) *Under Assumption 5.1, in the policy iteration algorithm, for each k we either have*

$$\vartheta_{k+1} < \vartheta_k$$

or else we have

$$\vartheta_{k+1} = \vartheta_k, \quad J_{k+1}(i) \leq J_k(i), \quad \textit{for } \ i = 1, 2, \ldots, n.$$

Furthermore the algorithm terminates and the policies μ_k and μ_{k+1} obtained upon termination are optimal.

□

In fact from the termination condition and Lemma 5.6 it follows that

$$\bar{T}_{\mu_k} J_k = \bar{T} J_k = \bar{T} J_{k+1} = \bar{T}_{\mu_{k+1}} J_{k+1}$$

upon termination. From Lemma 5.4 and Lemma 5.7, it is clear that $sp(J_k - J^\diamond) = 0$ and $\vartheta_k = \vartheta^*$ for $k \geq k^*$, where k^* is the step at which termination occurs and $J^\diamond$ is as in Equation 5.8.

Lemma 5.7 implies that the smallest k for which $\bar{T} J_k = \bar{T}_{\mu_k} J_k$ occurs is k^*. Proposition 5.2 and Lemma 5.7, also implies that for $l, k \leq k^*$, $\mu_l \neq \mu_k$ if $l \neq k$. Hence optimal policy is obtained within $|\Upsilon|$ steps, where Υ is the set of stationary deterministic policies and $|\Upsilon|$ is its cardinality.

Note that we could as well have imposed the termination condition to be: terminate if $\bar{T}_{\mu_k} J_k = \bar{T} J_k$.

Also note that at each step we could have used any differential cost J_k corresponding to policy μ_k (satisfying $\vartheta_k \underline{1} + J_k = \bar{T}_{\mu_k} J_k$), since this would not have changed policy μ_{k+1}. The termination condition will then be $\vartheta_k = \vartheta_{k+1}$ and $sp(J_k - J_{k+1}) = 0$.

Lemma 5.8 Let us consider a Unichain MDP. *Let μ and $\bar{\mu}$ be stationary deterministic policies such that $\bar{T}_{\bar{\mu}} J_\mu = \bar{T} J_\mu$, where J_μ is a differential cost for policy μ satisfying $\vartheta^\mu \underline{1} + J_\mu = \bar{T}_\mu J_\mu$. Suppose the Markov chain corresponding to $\bar{\mu}$ is*

irreducible. Then $\vartheta^{\bar{\mu}} < \vartheta^{\mu}$ if μ is not optimal and $\vartheta^{\bar{\mu}} = \vartheta^{\mu}$ if μ is optimal.

□

Proof of Lemma 5.8

Now $\vartheta^{\bar{\mu}} \leq \vartheta^{\mu}$ from Corollary 5.1. Of course if μ is optimal, $\vartheta^{\bar{\mu}}$ cannot be less than $\vartheta^{\mu} = \vartheta^*$. Hence $\vartheta^{\bar{\mu}} = \vartheta^{\mu}$. By Lemma 5.5

$$\vartheta^{\bar{\mu}}\underline{1} = \vartheta^{\mu}\underline{1} + P^*_{\bar{\mu}} \underbrace{(\bar{T}J_{\mu} - \bar{T}_{\mu}J_{\mu})}_{\leq \underline{0}}.$$

That is $\vartheta^{\bar{\mu}} = \vartheta^{\mu} + (\pi^{\bar{\mu}})'(\bar{T}J_{\mu} - \bar{T}_{\mu}J_{\mu})$. Here $\underline{0}$ is the zero vector and the inequality is componentwise. Note that $P^*_{\bar{\mu}} = \underline{1}(\pi^{\bar{\mu}})'$, where $\pi^{\bar{\mu}}$ is the unique invariant distribution corresponding to policy $\bar{\mu}$ (i.e. corresponding to transition probability matrix $P_{\bar{\mu}}$). $\pi^{\bar{\mu}}$ has all elements positive, since the Markov chain corresponding to policy $\bar{\mu}$ is irreducible.

Suppose μ is not optimal and $\vartheta^{\bar{\mu}} = \vartheta^{\mu}$. This implies $\bar{T}J_{\mu} = \bar{T}_{\mu}J_{\mu} = \vartheta^{\mu}\underline{1} + J_{\mu}$. Hence μ is optimal, a contradiction (see the comments following equation 5.7). Hence $\vartheta^{\bar{\mu}} < \vartheta^{\mu}$.

□

Lemma 5.8 says that $\vartheta^{\bar{\mu}} = \vartheta^{\mu}$ if and only if μ is optimal.

Now let us consider a recurrent MDP; where all stationary deterministic policies are irreducible. If we use policy iteration, $\vartheta^{\mu_{k+1}} < \vartheta^{\mu_k}$ if and only if μ_k is not optimal. Hence the termination condition could be $\vartheta_k = \vartheta_{k+1}$, at which stage μ_k and μ_{k+1} are optimal.

Lemma 5.9 *For a recurrent MDP, a stationary deterministic policy μ is optimal if and only if $\bar{T}J_\mu = \bar{T}_\mu J_\mu$; where J_μ is a differential cost for policy μ satisfying $\vartheta^\mu \underline{1} + J_\mu = \bar{T}_\mu J_\mu$.*

□

Proof of Lemma 5.9

The if part follows from equation 5.7 and the comments following that. Now to prove the only if part, note that if stationary deterministic policy $\bar{\mu}$ is such that, $\bar{T}_{\bar{\mu}} J_\mu = \bar{T} J_\mu$, then

$$\vartheta^{\bar{\mu}} = \vartheta^\mu + (\pi^{\bar{\mu}})' \underbrace{(\bar{T}J_\mu - \bar{T}_\mu J_\mu)}_{\leq \underline{0}}$$

by Lemma 5.5. $\pi^{\bar{\mu}}$ has all elements positive. Since μ is optimal $\vartheta^{\bar{\mu}} = \vartheta^\mu$. Hence $\bar{T}J_\mu = \bar{T}_\mu J_\mu$. □

We would like to add that, in general an MDP being unichain, does not mean that it has a common recurrent state. For instance consider the following example.

Example 5.1 *Consider the following three state (deterministic) MDP, where state space $\tilde{\mathcal{S}} = \{1, 2, 3\}$, and action set $\tilde{\mathcal{A}}(i) = \{1, 2\}$, $\forall i \in \tilde{\mathcal{S}}$. The transition probabilities are given by*

$$p_{i,(((i-1+u) \bmod 3)+1)}(u) = 1 \qquad \text{for } u \in \{1, 2\} = \tilde{\mathcal{A}}(i) \text{ and } i \in \{1, 2, 3\} = \tilde{\mathcal{S}}$$

There are eight different stationary deterministic policies, each of which are unichain, but has no common recurrent state.

□

5.5 Continuity Issues Of Limiting and Differential Matrices

In this section, let us define $\| P \|_\infty$ for a $\mathbf{R}^{n \times n}$ matrix P as

$$\| P \|_\infty \quad = \quad \max_{i,j} |p_{ij}| \tag{5.11}$$

where $p_{ij} = [P]_{ij}$. Note that this is a vector norm on $\mathbf{R}^{n \times n}$ matrices. Our first attempt might be to approximate any stationary deterministic policy with a stationary fully randomized policy; but this has the following problem, that for multichain policies the approximation by stationary fully randomized policies won't work as illustrated in the following example.

Example 5.2 *Let us consider a two state problem. With slight abuse of notation we use ϵ as the subscript for the matrices $P_\epsilon, \bar{P}_\epsilon, L_\epsilon$ and $\bar{L}_\epsilon$. Let*

$$\bar{P} = \begin{pmatrix} 1 & 0 \\ \\ 0 & 1 \end{pmatrix}, \quad P_\epsilon = \begin{pmatrix} 1-\epsilon & \epsilon \\ \\ \epsilon & 1-\epsilon \end{pmatrix}, \quad \bar{P}_\epsilon = \begin{pmatrix} 1-2\epsilon & 2\epsilon \\ \\ \epsilon & 1-\epsilon \end{pmatrix}.$$

for $0 < \epsilon < \frac{1}{2}$. Note that

$$\left.\begin{array}{rcl} \lim_{\epsilon \downarrow 0} \| P_\epsilon - \bar{P} \|_\infty & = & 0 \\ \lim_{\epsilon \downarrow 0} \| \bar{P}_\epsilon - \bar{P} \|_\infty & = & 0 \\ \lim_{\epsilon \downarrow 0} \| \bar{P}_\epsilon - P_\epsilon \|_\infty & = & 0 \end{array}\right\} \tag{5.12}$$

Note that we have for $\epsilon \in (0, \frac{1}{2})$

$$\begin{aligned} P^*_\epsilon \quad &\equiv \quad \lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N} P^k_\epsilon \quad = \quad \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}, \\ L_\epsilon \quad &\equiv \quad (I - P_\epsilon + P^*_\epsilon)^{-1} - P^*_\epsilon \end{aligned}$$

$$= \begin{pmatrix} \frac{1}{4\epsilon} & -\frac{1}{4\epsilon} \\ \\ -\frac{1}{4\epsilon} & \frac{1}{4\epsilon} \end{pmatrix},$$

$$\bar{P}^*_\epsilon \equiv \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N} \bar{P}^k_\epsilon = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} \\ \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix},$$

$$\bar{L}_\epsilon \equiv (I - \bar{P}_\epsilon + \bar{P}^*_\epsilon)^{-1} - \bar{P}^*_\epsilon$$

$$= \begin{pmatrix} \frac{2}{9\epsilon} & -\frac{2}{9\epsilon} \\ \\ -\frac{1}{9\epsilon} & \frac{1}{9\epsilon} \end{pmatrix},$$

$$\bar{P}^* \equiv \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N} \bar{P}^k = \begin{pmatrix} 1 & 0 \\ \\ 0 & 1 \end{pmatrix},$$

$$\bar{L} \equiv (I - \bar{P} + \bar{P}^*)^{-1} - \bar{P}^*$$

$$= \begin{pmatrix} 0 & 0 \\ \\ 0 & 0 \end{pmatrix}.$$

Note that in spite of the relation 5.12,

$$\lim_{\epsilon\downarrow 0} \| P^*_\epsilon - \bar{P}^* \|_\infty \neq 0, \quad \lim_{\epsilon\downarrow 0} \| L_\epsilon - \bar{L} \|_\infty = \infty \neq 0.$$

$$\lim_{\epsilon\downarrow 0} \| \bar{P}^*_\epsilon - \bar{P}^* \|_\infty \neq 0, \quad \lim_{\epsilon\downarrow 0} \| \bar{L}_\epsilon - \bar{L} \|_\infty = \infty \neq 0.$$

$$\lim_{\epsilon\downarrow 0} \| P^*_\epsilon - \bar{P}^*_\epsilon \|_\infty \neq 0, \quad \lim_{\epsilon\downarrow 0} \| L_\epsilon - \bar{L}_\epsilon \|_\infty = \infty \neq 0.$$

Consider the two state deterministic problem, where state space $\tilde{\mathcal{S}} = \{1, 2\}$ *and action space* $\tilde{\mathcal{A}}(i) = \{1, 2\}$ *for* $i \in \mathcal{S}$. *Also* $g(1,1) = g(1,2) = 1$ *and* $g(2,1) = g(2,2) = 0$. *Let* $p_{11}(1) = 1,\ p_{12}(2) = 1,\ p_{21}(1) = 1,\ p_{22}(2) = 1$. *Let* μ *be a stationary deterministic policy such that* $\mu(1) = 1,\ \mu(2) = 2$. *For* $0 < \epsilon < \frac{1}{2}$, *let* δ_ϵ *and* $\bar{\delta}_\epsilon$ *be stationary fully randomized policies (unichain) such that*

$$[\delta_\epsilon(1)]_1 = 1 - \epsilon,\ [\delta_\epsilon(1)]_2 = \epsilon,\ [\delta_\epsilon(2)]_1 = \epsilon,\ [\delta_\epsilon(2)]_2 = 1 - \epsilon.$$

and

$$[\bar{\delta}_\epsilon(1)]_1 = 1 - 2\epsilon,\ [\bar{\delta}_\epsilon(1)]_2 = 2\epsilon,\ [\bar{\delta}_\epsilon(2)]_1 = \epsilon,\ [\bar{\delta}_\epsilon(2)]_2 = 1 - \epsilon.$$

Now the immediate cost vectors are given by

$$\bar{g}^\mu = \begin{pmatrix} 1 \\ \\ 0 \end{pmatrix}, \quad \bar{g}^{\delta_\epsilon} = \begin{pmatrix} 1 \\ \\ 0 \end{pmatrix}, \quad \bar{g}^{\bar{\delta}_\epsilon} = \begin{pmatrix} 1 \\ \\ 0 \end{pmatrix}.$$

Also $P_\mu = \bar{P}$, $P_{\delta_\epsilon} = P_\epsilon$ *and* $P_{\bar{\delta}_\epsilon} = \bar{P}_\epsilon$. *The average cost vectors are*

$$\begin{aligned} \bar{\vartheta}^\mu &= P^*_\mu \bar{g}^\mu = \begin{pmatrix} 1 \\ \\ 0 \end{pmatrix}, \\ \bar{\vartheta}^{\delta_\epsilon} &= P^*_{\delta_\epsilon} \bar{g}^{\delta_\epsilon} = \begin{pmatrix} \frac{1}{2} \\ \\ \frac{1}{2} \end{pmatrix}, \\ \bar{\vartheta}^{\bar{\delta}_\epsilon} &= P^*_{\bar{\delta}_\epsilon} \bar{g}^{\bar{\delta}_\epsilon} = \begin{pmatrix} \frac{1}{3} \\ \\ \frac{1}{3} \end{pmatrix}. \end{aligned}$$

Consider the differential cost vectors

$$\begin{aligned} J_{\delta_\epsilon} &= L_{\delta_\epsilon} \bar{g}^{\delta_\epsilon} = L_\epsilon \bar{g}^{\delta_\epsilon} = \begin{pmatrix} \frac{1}{4\epsilon} \\ \\ -\frac{1}{4\epsilon} \end{pmatrix}, \\ J_{\bar{\delta}_\epsilon} &= L_{\bar{\delta}_\epsilon} \bar{g}^{\bar{\delta}_\epsilon} = \bar{L}_\epsilon \bar{g}^{\bar{\delta}_\epsilon} = \begin{pmatrix} \frac{2}{9\epsilon} \\ \\ -\frac{1}{9\epsilon} \end{pmatrix}. \end{aligned}$$

Note that though P_{δ_ϵ} *and* $P_{\bar{\delta}_\epsilon}$ *converge to* P_μ *in the* $\|\cdot\|_\infty$ *norm,*

$$\begin{aligned} \lim_{\epsilon \downarrow 0} \| \bar{\vartheta}^{\delta_\epsilon} - \bar{\vartheta}^\mu \|_\infty &\neq 0, \\ \lim_{\epsilon \downarrow 0} \| \bar{\vartheta}^{\bar{\delta}_\epsilon} - \bar{\vartheta}^\mu \|_\infty &\neq 0, \\ \lim_{\epsilon \downarrow 0} \| \bar{\vartheta}^{\delta_\epsilon} - \bar{\vartheta}^{\bar{\delta}_\epsilon} \|_\infty &\neq 0. \end{aligned}$$

where for a vector $J \in \mathbf{R}^n$,

$$\| J \|_\infty = \max_i |J(i)|.$$

Also

$$sp(J_{\delta_\epsilon} - J_{\bar{\delta}_\epsilon}) \;=\; \begin{pmatrix} \frac{1}{36\epsilon} \\ \\ -\frac{5}{36\epsilon} \end{pmatrix} \;=\; \frac{1}{6\epsilon}.$$

Note that

$$\lim_{\epsilon \downarrow 0} sp(J_{\delta_\epsilon} - J_{\bar{\delta}_\epsilon}) = \infty \;\neq\; 0.$$

□

But however we have the following lemma [40, Proposition 8.4.6]

Lemma 5.10 *Let* $\{P_k : k \geq 0\}$ *denote a sequence of unichain transition probability (or stochastic) matrices and suppose*

$$\lim_{k \to \infty} \| P_k - P \|_\infty = 0,$$

where P *is also a unichain stochastic matrix. Then*
(a)

$$\lim_{k \to \infty} \| P_k^* - P^* \|_\infty = 0,$$

(b)

$$\lim_{k \to \infty} \| L_k - L \|_\infty = 0,$$

where P_k^* *and* L_k *are the limiting matrix and differential matrix corresponding to* P_k *(see the statements following Lemma 5.1). Similarly* P^* *and* L *are the limiting and differential matrix, respectively of the stochastic matrix* P. □

Note that in [40, Proposition 8.4.6] we should impose the condition that P is unichain; else the result does not hold, as was shown in the previous example.

Note that, by the results in Appendix B, for a unichain MDP, if μ is a stationary deterministic policy and δ is a stationary randomized policy that *subsumes* policy μ (that is $[\delta(i)]_{\mu(i)} > 0$ for $i \in \{1, 2, \ldots, n\}$), then δ is also unichain and has as its recurrent class, a super set of the unique recurrent class of μ.

Hence for a unichain MDP, if $\{\delta_n\}$ is a sequence of stationary randomized policies which "converge" to a stationary deterministic policy μ, Then the immediate cost vectors

$$\bar{g}^{\delta_k} \underset{k\to\infty}{\longrightarrow} \bar{g}^{\mu}$$

and the average cost (same for all starting states)

$$\vartheta^{\delta_k} \underset{k\to\infty}{\longrightarrow} \vartheta^{\mu}$$

Also the basic differential cost vectors

$$J^*_{\delta_k} \underset{k\to\infty}{\longrightarrow} J^*_{\mu}$$

where $J^*_{\delta_k} = L_{\delta_k}\bar{g}^{\delta_k}$ and $J^*_{\mu} = L_{\mu}\bar{g}^{\mu}$. Here L_{δ_k} and L_{μ} are the differential matrices corresponding to P_{δ_k} and P_{μ} (the stochastic matrices corresponding to policies δ_k and μ respectively).

Thus we can approximate stationary deterministic policies with stationary randomized policies.

Lemma 5.11 *Let P be a stochastic matrix corresponding to a unichain MDP. Then there exists an $\epsilon > 0$, such that if $\bar{P}$ is any stochastic matrix with $\| P - \bar{P} \|_\infty \leq \epsilon$,*

then the Markov chain corresponding to $\bar{P}$ is also unichain, with the recurrent class of $\bar{P}$ being a super-set of the recurrent class of P.

□

Proof of Lemma 5.11

Choose

$$\epsilon = \frac{1}{2}\left(\min_{i,j \text{ with } p_{ij}\neq 0} p_{ij}\right)$$

Hence $p_{ij} > 0$ implies $\bar{p}_{ij} > 0$. Here $\bar{p}_{ij} = [\bar{P}]_{ij}$. It is easily seen that any recurrent state under P is recurrent under $\bar{P}$ (since any recurrent state under P is reachable or accessible with positive probability from any state under $\bar{P}$).

□

5.6 Approximate Policy Iteration

In the following $\|\cdot\|$ denotes the sup-norm. We have the following lemma.

Lemma 5.12 *Let $\tilde{J}$ be any fixed vector in $\mathbf{R}^n$. Then there exist scalars $\varsigma > 0, \epsilon > 0$, depending on $\tilde{J}$ such that if J is any vector in $\mathbf{R}^n$, with $sp(J - \tilde{J}) < \epsilon$ and $\tilde{\mu}$ is a stationary deterministic policy such that*

$$\| \bar{T}_{\tilde{\mu}} J - \bar{T} J \| < \varsigma$$

then

$$\bar{T}_{\tilde{\mu}} \tilde{J} = \bar{T} \tilde{J}$$

□

Proof of Lemma 5.12

This follows from the affine nature of the operator $\bar{T}_\mu$ for any stationary deterministic policy μ along with the property that $\bar{T}_\mu(V + \alpha\underline{1}) = \bar{T}_\mu V + \alpha\underline{1}$ (where $V \in \mathbf{R}^n,\ \alpha \in \mathbf{R}$) and monotonicity of $\bar{T}_\mu$ (i.e. if $V, \tilde{V} \in \mathbf{R}^n$ and $V \geq \tilde{V}$ then $\bar{T}_\mu V \geq \bar{T}_\mu \tilde{V}$. Here the inequality is componentwise). Also $\bar{T}(V + \alpha\underline{1}) = \bar{T}V + \alpha\underline{1}$ and $\bar{T}$ is a monotone operator. Also note that the set of stationary deterministic policies is finite (i.e. $|\Upsilon|$ is finite). □

Corollary 5.2 *For any finite state, finite action unichain MDP, there exist scalars $\epsilon > 0$, $\varsigma > 0$, such that if J is any vector in $\mathbf{R}^n$ with $sp(J - J_\mu) < \epsilon$ (where J_μ is a differential cost vector for stationary deterministic policy μ) and $\tilde{\mu}$ is a stationary deterministic policy such that*

$$\| \bar{T}_{\tilde{\mu}} J - \bar{T} J \| < \varsigma$$

then $\bar{T}_{\tilde{\mu}} J_\mu = \bar{T} J_\mu$. In fact because $|\Upsilon|$ is finite, the ϵ and ς can be chosen to be uniformly applicable to all $\mu \in \Upsilon$. □

Consider a finite state, finite action unichain MDP with a common recurrent state. Assume that a sequence of stationary deterministic policies $\{\mu_k\}$ and a corresponding sequence of approximate differential cost vectors $\{J_k\}$ satisfy

$$sp(J_k - J_{\mu_k}) \leq \epsilon_k \qquad \text{for } k = 0, 1, 2, \ldots$$

and

$$\| \bar{T}_{\mu_{k+1}} J_k - \bar{T} J_k \| \leq \varsigma_k \qquad \text{for } k = 0, 1, 2, \ldots$$

where J_{μ_k} is a differential cost for stationary deterministic policy μ_k.

Then there exists an $\epsilon > 0, \varsigma > 0$, such that if $\epsilon_k \leq \epsilon, \varsigma_k \leq \varsigma$, $\forall k$, then the sequence of policies μ_k generated are the same as that would be generated in the policy iteration scheme and hence ϑ^{μ_k} converges to ϑ^* (the optimal average cost) in a finite ($\leq |\Upsilon|$) number of steps.

Observe that if $\limsup_{k\to\infty} \epsilon_k < \epsilon$ and $\limsup_{k\to\infty} \varsigma_k < \varsigma$, then $\vartheta^{\mu_k} = \vartheta^*$ and $sp(J_{\mu_k} - J^\diamond) = 0$ for all large k. Here $J^\diamond$ is the unique vector satisfying

$$\vartheta^* \underline{1} + J^\diamond = \bar{T} J^\diamond$$

with $J^\diamond(k) = 0$ for some fixed state $k \in \mathcal{S}$.

For error bounds for the average cost problem, see Section C.6 in Appendix C. These results can be used to check the nearness of convergence of the approximate policy iteration schemes developed in Section 5.9.

5.7 Average Cost Temporal Difference Schemes

The purpose of the present section is to discuss a variant of TD (Temporal Difference) learning that is suitable for approximating differential cost functions of undiscounted (average cost) Markov chains (i.e. solutions to Poisson's equation) [55]. Actually we are dealing with Markov Cost Processes. The results parallel those available for the discounted cost (see [54] and Chapter 4) : we have convergence with probability one, a characterization of the limit , and graceful bounds on the resulting approximation error. Note that [55] discusses only a finite state irreducible aperiodic Markov chain. We extend this to any finite state irreducible Markov chain

(periodic or otherwise) where the immediate cost may be random but stationary. The results are essentially the same, and the proof is almost on the same lines as in [55], but one of the main differences is in the use of results from Chapter 3. Additional variations in the proofs of sub-results leading to the main results are given as and when required.

Consider the homogeneous Markov chain with the state space $\mathcal{S} = \{1, 2, \ldots, n\}$ and the $n \times n$ transition probability matrix $P = [p_{ij}]$. Let g_t denote the immediate cost incurred while making a transition from state i_t at time t to state i_{t+1} at time $t+1$; the cost may be random but has finite variance and mean. Let $g(i)$ denote the expected value of the immediate cost incurred from state i, namely $E[g_t \mid i_t = i]$. The probability distribution of g_t may depend on i_t and i_{t+1}, but given i_t and i_{t+1}, does not depend on the past values of i_l and g_l $(l < t)$ (Markov property).

Assumption 5.2 *The Markov chain corresponding to P is irreducible (may be aperiodic or periodic).* □

It follows that the Markov chain has a unique invariant probability distribution $\pi \in \mathbf{R}^n$, that satisfies $\pi' P = \pi'$, with $\pi(i) > 0$ for all $i \in \mathcal{S}$. Let $E_0[\cdot]$ denote expectation with respect to this distribution. We define the average cost by

$$\vartheta^* = E_0[g_t] = \sum_{i=1}^{n} \pi(i) g(i),$$

and a differential cost function as any function $J : \mathcal{S} \to \mathbf{R}$, satisfying *Poisson's*

equation which takes the form

$$J = \bar{g} - \vartheta^* \underline{1} + PJ$$

Here $\bar{g} \in \mathbf{R}^n$ is the expected immediate cost column vector, whose i^{th} component is $g(i)$. J is viewed as a vector in $\mathbf{R}^n$.

Under Assumption 5.2, it is known that the differential cost functions exist and the set of all differential cost functions takes the form $\{J^* + c\underline{1} | c \in \mathbf{R}\}$ for some function satisfying $\pi' J^* = 0$ (see Lemma 5.3 and the statements after that). We will refer J^* as the *basic* differential cost function and it is known that under Assumption 5.2,

$$J^* = L\bar{g} = C - \lim_{N \to \infty} \sum_{t=0}^{N} P^t (\bar{g} - \vartheta^* \underline{1}), \tag{5.13}$$

where

$$L = (I - P + P^*)^{-1} - P^*,$$

is the differential matrix corresponding to P. Here

$$P^* = \lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} P^k = \underline{1}\pi'$$

is the limiting matrix corresponding to P.

Neither P nor the distribution of the immediate cost g_t is known in advance. In a general setting of the TD(λ) scheme [55] we consider approximations to differential cost functions using a function of the form

$$\tilde{J}(i, r) = \sum_{k=1}^{K} r(k) \phi_k(i)$$

where $r = (r(1), r(2), \ldots, r(K))'$ is a tunable parameter vector and each $\phi_k \in \mathbf{R}^n$ is a basis function defined on the state space $\mathcal{S}$.

It is convenient to define a vector valued function $\phi : \mathcal{S} \to \mathbf{R}^K$, by letting

$$\phi(i) = (\phi_1(i), \phi_2(i), \ldots, \phi_K(i))'.$$

With this notation, the approximation can also be written in the form $\tilde{J}(i,r) = r'\phi(i)$ or $\tilde{J} = \Phi r$, where Φ is an $n \times K$ matrix whose k^{th} column is equal to ϕ_k.

Assumption 5.3

(a) *The basis functions* $\{\phi_k | k = 1, \ldots, K\}$ *are linearly independent; i.e.* Φ *has full column rank.*

(b) *For every* $r \in \mathbf{R}^K$, $\Phi r \neq \underline{1}$, *i.e.* $\underline{1}$ *does not lie in the space spanned by* $\phi_k s$, $k = 1, 2, \ldots, K$.

□

In particular $K < n$.

Suppose that we observe a sequence of states i_t generated according to the transition probability matrix P. Given that at time t, the parameter vector r has been set to some value r_t, and we have an approximation ϑ_t to the average cost ϑ^*, we define the temporal difference d_t corresponding to the transition from state i_t to state i_{t+1} by

$$d_t \;=\; g_t - \vartheta_t + \tilde{J}(i_{t+1}, r_t) - \tilde{J}(i_t, r_t) \tag{5.14}$$

The TD(λ) algorithm updates r_t and ϑ_t according to

$$\vartheta_{t+1} \;=\; (1 - \eta_t)\vartheta_t + \eta_t g_t \tag{5.15}$$

and

$$r_{t+1} = r_t + \gamma_t d_t \sum_{k=0}^{t} \lambda^{t-k} \phi(i_k) \tag{5.16}$$

where γ_t and η_t are scalar step sizes and λ is a parameter in $[0, 1)$. Define eligibility vectors $z_t \in \mathbf{R}^K$, by

$$z_t = \sum_{k=0}^{t} \lambda^{t-k} \phi(i_k) \tag{5.17}$$

With this new notation, the parameter updates are given by

$$\begin{aligned} r_{t+1} &= r_t + \gamma_t d_t z_t \\ z_{t+1} &= \lambda z_t + \phi(i_{t+1}) \end{aligned}$$

with $z_{-1} = 0$.

Assumption 5.4

(a) *The non-negative monotonic step sizes $\gamma_t \downarrow 0$ are pre-determined (deterministic) and satisfy*

$$\sum_{t=0}^{\infty} \gamma_t = \infty; \qquad \sum_{t=0}^{\infty} \gamma_t^2 < \infty.$$

(b) *There exists a positive scalar c such that the sequence η_t satisfies $\eta_t = c\gamma_t$ for all $t \geq 0$.*

□

Note the variation in Assumption 5.4(a) from [55, Assumption 3(a)]. Actually Assumption 5.4(a) may be replaced by (see Chapter 3)

5.4(a') The non-negative step sizes γ_t are pre-determined and satisfy

$$\sum_{t=0}^{\infty} \gamma_t = \infty, \quad \sum_{t=0}^{\infty} \gamma_t^2 < \infty, \quad \text{and} \quad \sum_{t=0}^{\infty} |\gamma_{t+1} - \gamma_t| < \infty.$$

Assumption 5.5 *The immediate cost g_t has finite moments, i.e.*

$$E\left[|g_t|^k \mid i_t = i\right] < \infty; \quad \forall i \in \mathcal{S},\ \forall k \in \mathbf{N}$$

where $\mathbf{N}$ *is the set of natural numbers.*

□

5.7.1 Convergence Results

We define an $n \times n$ diagonal matrix D with diagonal entries $\pi(1), \pi(2), \ldots, \pi(n)$. It is easy to see that $< x, y >_D \equiv x'Dy,\ x, y \in \mathbf{R}^n$ defines an inner product space, with norm

$$\| x \|_D = \sqrt{< x, x >_D}$$

We say that two vectors J and $\bar{J}$ are D-orthogonal if $J'D\bar{J} = 0$. Here J' is the transpose of the vector J. In this section we use $\| \cdot \|$ without a subscript, to denote the Eucledian norm on vectors or the Eucledian induced norm on matrices (that is for any matrix M, we have $\| M \| = \max_{\|x\|=1} \| Mx \|$.)

We define the projection matrix Π that projects onto the subspace spanned by the basis functions. In particular $\Pi = \Phi(\Phi'D\Phi)^{-1}\Phi'D$. For any $J \in \mathrm{R}^n$, we then have

$$\Pi J = \arg \min_{\bar{J} \in \{\Phi r | r \in \mathbf{R}^K\}} \| J - \bar{J} \|_D$$

Note that

$$\Phi'D(J - \Pi J) = \underline{0}$$

where $\underline{0}$ is a K dimensional zero vector.

In fact ΠJ is the unique vector which lies in the span of ϕ_ks ($k = 1, 2, \ldots, K$) such that $(J - \Pi J)$ is D-orthogonal to all ϕ_ks ($k = 1, 2, \ldots, K$).

For any $\lambda \in [0, 1)$, we define an operator $\bar{T}^{(\lambda)} : \mathbf{R}^n \to \mathbf{R}^n$ by

$$\bar{T}^{(\lambda)} J = (1 - \lambda) \sum_{m=0}^{\infty} \lambda^m \left(\sum_{t=0}^{m} P^t (\bar{g} - \vartheta^* \underline{1}) + P^{m+1} J \right),$$

where $J \in \mathbf{R}^n$. Note that since the elements of the vector $\sum_{t=0}^{m} P^t (\bar{g} - \mu^* \underline{1})$ grows at most linearly in m, the outer sum is well defined. In fact it may be shown that the elements of the vector $\sum_{t=0}^{m} P^t (\bar{g} - \mu^* \underline{1})$ is bounded. Thus $\bar{T}^{(\lambda)}$ is an affine function. From the relation

$$(I - \lambda P)^{-1} \;=\; \sum_{m=0}^{\infty} (\lambda P)^m \;=\; (1 - \lambda) \sum_{m=0}^{\infty} \lambda^m \sum_{t=0}^{m} P^t, \tag{5.18}$$

we may rewrite

$$\bar{T}^{(\lambda)} J \;=\; (I - \lambda P)^{-1} (\bar{g} - \vartheta^* \underline{1}) + P^{(\lambda)} J$$

where P^{λ} is defined later in equation 5.20. Our convergence result follows [55, Theorem 1].

Theorem 5.1 *Under Assumptions 5.2–5.5, the following hold:*

(a) *For any $\lambda \in [0, 1)$, the average cost* TD*(λ) algorithm, as defined in the earlier part of this section converges with probability one.*

(b) *The limit of convergence of the sequence ϑ_t is the average cost ϑ^*.*

(c) *The limit r^* of the sequence r_t is the unique solution of the equation*

$$\Pi \bar{T}^{(\lambda)} (\Phi r^*) = \Phi r^*.$$

□

We follow along the lines of [55].

5.7.1.1 Preliminaries

Construct a process $X_t = (i_t, i_{t+1}, z_t, g_t)$, where z_t is the eligibility vector z_t defined in equation 5.17. It is easy to see that X_t is a Markov process. In particular z_{t+1} and i_{t+1} are deterministic functions of X_t, and the distribution of i_{t+2} only depends on i_{t+1}; also the distribution of g_{t+1} depends only on i_{t+1} and i_{t+2}. Note that at each time t, the random vector X_t, together with the current values of ϑ_t and r_t, provides all necessary information for computing ϑ_{t+1} and r_{t+1}.

So that we can think of the TD(λ) algorithm as adapting only a single vector, we introduce a sequence $\theta_t \in \mathbf{R}^{K+1}$ with components, $\theta_t(1) = \vartheta_t$ and $\theta_t(i) = r_t(i-1)$ for $i \in \{2, \ldots, K+1\}$, or using more compact notation,

$$\theta_t = \begin{bmatrix} \vartheta_t \\ \\ r_t \end{bmatrix}$$

The TD(λ) updates can be rewritten as

$$\theta_{t+1} = \theta_t + \gamma_t(A(X_t)\theta_t + b(X_t)), \tag{5.19}$$

where for any $X = (i, j, z, g)$, we have

$$A(X) = \begin{bmatrix} -c & 0 \cdots 0 \\ \\ -z & z(\phi'(j) - \phi'(i)) \end{bmatrix}$$

and

$$b(X) = \begin{bmatrix} cg \\ \\ zg \end{bmatrix}$$

and c is the constant in Assumption 5.4(b). As is shown in [55], $A(X_t)$ and $b(X_t)$ have well known "steady state" expectation which we denote by A and b. Note that [55] deals with the case where g_t depends only on i_t; but it easily extends to the case where g_t is random.

General results concerning stochastic approximation algorithms can be used to show that the asymptotic behaviour of the sequence generated by equation 5.19, mimics that of an ordinary differential equation

$$\dot{\theta}_t = A\theta_t + b.$$

We essentially use the very general stochastic approximation result (Theorem 3.2, Chapter 3) to prove that θ_t converges with probability one.

5.7.1.2 Lemmas

Lemma 5.13 *Under Assumption 5.2, for all $J \in \mathbf{R}^n$,*

$$\| PJ \|_D \leq \| J \|_D .$$

Furthermore, unless J is proportional to $\underline{1}$, we have $PJ \neq J$. □

Proof of Lemma 5.13

The first part of the lemma is proved in [54]. We prove the second part as follows. If J is proportional to $\underline{1}$, it is easy to see that $PJ = J$.

Suppose $PJ = J$. This implies $P^2J = PJ = J$. Continuing similarly $P^kJ = J$, $\forall k \geq 1$. Hence

$$\sum_{k=0}^{N-1} P^k J = NJ,$$

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} P^k J = J.$$

That is

$$P^* J = J,$$

$$\underline{1}\pi' J = J.$$

This implies that J is proportional to $\underline{1}$. □

Under Assumption 5.2, the matrix $P^{(\lambda)}$ defined below in equation 5.20, is an irreducible stochastic matrix for $\lambda \in [0, 1)$. Note that $P^{(0)} = P$. Furthermore for $\lambda \in (0, 1)$, $P^{(\lambda)}$ is aperiodic (actually all elements of $P^{(\lambda)}$ are positive) even when P is periodic. Also $P^{(\lambda)}$ has unique invariant distribution π.

Lemma 5.14 *Let*

$$P^{(\lambda)} = (1-\lambda) \sum_{m=0}^{\infty} \lambda^m P^{m+1} \tag{5.20}$$

Then under Assumption 5.2, for any $\lambda \in [0, 1)$ and $J \in \mathbf{R}^n$,

$$\| P^{(\lambda)} J \|_D \leq \| J \|_D$$

Furthermore unless J is proportional to $\underline{1}$, we have $P^{(\lambda)} J \neq J$.

□

Proof of Lemma 5.14 is similar to that of Lemma 5.13. Note that $P^{(\lambda)}$ is continuous at $\lambda = 0$. Also for $\lambda \in (0, 1)$

$$P^{(\lambda)} = \frac{1-\lambda}{\lambda}((I - \lambda P)^{-1} - I).$$

Hence $P^{(\lambda)}$ is continuous in $\lambda \in [0,1)$. The proof of the following lemma needs a different line of argument than in [55] for the general (not necessarily aperiodic) case.

Lemma 5.15 *Under Assumption 5.2, for any $\lambda \in [0,1)$, we have $\bar{T}^{(\lambda)}J = J$ if and only if $J \in \{J^* + c\underline{1} | c \in \mathbf{R}\}$.*

□

Proof of Lemma 5.15

By Lemma 5.1 in the earlier section

$$\begin{aligned} P^* + L &= I + PL, \\ P^*P &= PP^* = P^*, \end{aligned}$$

where

$$\begin{aligned} P^* &= \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} P^k, \\ L &= (I - (P - P^*))^{-1} - P^*. \end{aligned}$$

Also

$$J^* = L\bar{g}.$$

Note that

$$\begin{aligned} PL &= P^* + L - I, \\ P^2L &= P^* + PL - P \\ &= 2P^* + L - (I + P). \end{aligned}$$

By induction for $k \geq 1$

$$P^kL = kP^* + L - (\sum_{l=0}^{k-1} P^l),$$

where $P^0 = I$.

Suppose $J = J^* + c\underline{1}$ for some scalar c. Then

$$\begin{aligned}
&\bar{T}^{(\lambda)} J \\
&= (1-\lambda) \sum_{m=0}^{\infty} \lambda^m \left(\sum_{t=0}^{m} P^t (\bar{g} - \vartheta^* \underline{1}) + P^{m+1} (J^* + c\underline{1}) \right) \\
&= (1-\lambda) \sum_{m=0}^{\infty} \lambda^m \left(\sum_{t=0}^{m} P^t \bar{g} - \sum_{t=0}^{m} P^t P^* \bar{g} + P^{m+1} L \bar{g} \right) + c\underline{1} \\
&= (1-\lambda) \sum_{m=0}^{\infty} \lambda^m \left(\sum_{t=0}^{m} P^t \bar{g} - (m+1) P^* \bar{g} + (m+1) P^* \bar{g} + L\bar{g} - \sum_{l=0}^{m} P^l \bar{g} \right) + c\underline{1} \\
&= (1-\lambda) \sum_{m=0}^{\infty} \lambda^m L \bar{g} + c\underline{1} \\
&= L\bar{g} + c\underline{1} \\
&= J^* + c\underline{1} \\
&= J.
\end{aligned}$$

The only if part of the proof is as in [55], which we include for completeness. Suppose J is not of the form $J^* + c\underline{1}$. Then

$$\begin{aligned}
\bar{T}^{(\lambda)} J &= \bar{T}^{(\lambda)} J^* + P^{(\lambda)} (J - J^*) \\
&= J^* + P^{(\lambda)} (J - J^*) \\
&\neq J^* + (J - J^*) \\
&= J,
\end{aligned}$$

where the inequality follows from Lemma 5.14. □

The process X_t constructed earlier is a Markov process with a steady state behaviour. Let $\mathcal{X}$ be the state space for the process. Let $E_0[\cdot]$ denote the expectation with respect to the invariant distribution of this process [55]. An argument along

the same line as in [55], gives the following lemma. See also [54].

Lemma 5.16 *Under Assumption 5.2, the steady state invariant expectations* $A = E_0[A(X_t)]$ *and* $b = E_0[b(X_t)]$ *are given by*

$$A = \begin{bmatrix} -c & 0 \cdots 0 \\ -\frac{1}{1-\lambda}\Phi' D \underline{1} & \Phi' D(P^{(\lambda)} - I)\Phi \end{bmatrix},$$

and

$$b = \begin{bmatrix} c\vartheta^* \\ \Phi' D(1-\lambda)\sum_{m=0}^{\infty}\lambda^m \sum_{t=0}^{m} P^t \bar{g} \end{bmatrix}.$$

□

Consider the following Markov chain derived from the original irreducible Markov chain with state space $\mathcal{S}$; the state space being

$$\tilde{\mathcal{S}} = \{(i,j) | i,j \in \mathcal{S};\ p_{ij} > 0\}.$$

The transition probability is defined as

$$\Pr\{(i_{t+1}, j_{t+1}) = (i', j') | (i_t, j_t) = (i,j)\} = p_{i'j'} \boldsymbol{\mathcal{I}}_{[j=i']}$$

for $(i,j) \in \tilde{\mathcal{S}}$, $(i',j') \in \tilde{\mathcal{S}}$; where the indicator function $\boldsymbol{\mathcal{I}}_{[j=i']} = 1$, if $i' = j$, else equal to zero.

It may be seen that this new Markov chain with state space $\tilde{\mathcal{S}}$ is irreducible and has period 'd', the same period as that for the original Markov chain with state space $\mathcal{S}$. The state space $\tilde{\mathcal{S}}$ may be partitioned into periodic classes $\tilde{\mathcal{C}}_0, \ldots, \tilde{\mathcal{C}}_{d-1}$; such that

$$\bigcup_{l=0}^{d-1} \tilde{\mathcal{C}}_l = \tilde{\mathcal{S}}; \quad \tilde{\mathcal{C}}_i \cap \tilde{\mathcal{C}}_j = \emptyset \text{ for } i \neq j.$$

Also

$$\Pr\{(i_{t+1}, j_{t+1}) \in \tilde{\mathcal{C}}_{((l+1) \bmod d)} | (i_t, j_t) = (i, j)\} = 1$$

for all $(i, j) \in \tilde{\mathcal{C}}_l$, $l \in \{0, \ldots, d-1\}$. Hence we can partition the state space $\mathcal{X}$ into disjoint Borel sets $\mathcal{X}_0, \mathcal{X}_1, \ldots, \mathcal{X}_{d-1}$ with

$$\Pi_\theta(x, \mathcal{X}_{(l+1) \bmod d}) = 1, \quad \forall x \in \mathcal{X}_l$$

where Π_θ is the transition probability kernel for the Markov process X_t, when $\theta \in \mathbf{R}^{K+1}$. (Note the slight abuse of notation in the use of Π_θ for the transition probability kernel and Π for the projection matrix). Note that in our case Π_θ is independent of θ. Note that

$$\mathcal{X}_l = \{(i, j, z, g) | (i, j) \in \tilde{\mathcal{C}}_l, z \in \mathbf{R}^K, g \in \mathbf{R}\}.$$

We are a bit imprecise in the definition of $\mathcal{X}_l$ in that, actually z might take values in a proper sub-set of $\mathbf{R}^K$, which is dependent on (i, j), λ and the choice of ϕ_ks. Similarly g might take values in a proper sub-set of $\mathbf{R}$, which is dependent on (i, j). We say that a square matrix $M \in \mathbf{R}^{n \times n}$ is negative definite if $x'Mx < 0$ for all $x \in \mathbf{R}^n$, $x \neq \underline{0}$; even if M is not symmetric. The matrix A is not necessarily negative definite, but becomes negative definite under an appropriate co-ordinate scaling.

Lemma 5.17 *Under Assumption 5.2 and Assumption 5.3, $\Phi' D(P^{(\lambda)} - I)\Phi$ is negative definite.*

□

Proof of Lemma 5.17

Let J be a non-constant function on the state space $\mathcal{S}$. Since the Markov chain $\{i_t\}$ is irreducible, $J(i_t)$ is not a constant function of time, which implies that

$$\begin{aligned} 0 &< \frac{1}{2}E_0\left[(J(i_{t+1}) - J(i_t))^2\right] \\ &= E_0\left[J^2(i_t)\right] - E_0\left[J(i_t)J(i_{t+1})\right] \\ &= J'DJ - J'DPJ \\ &= J'D(I-P)J \end{aligned} \tag{5.21}$$

For any $r \neq 0$, $J = \Phi r$, is a nonconstant vector because of Assumption 5.3. Thus $r'\Phi'D(P-I)\Phi r < 0$, for every $r \neq \underline{0}$, which shows that the matrix $\Phi'D(P-I)\Phi$ is negative definite. The same argument works for the matrix $\Phi'D(P^{(\lambda)}-I)\Phi$, because $P^{(\lambda)}$ is also an irreducible stochastic matrix with the same invariant distribution.

□

Another way for deriving equation 5.21, is as follows. Since the Markov chain is irreducible and hence $\pi(i) > 0$, $\forall i \in \mathcal{S}$ and also since J is a non-constant function

$$\begin{aligned} 0 &< \frac{1}{2}E_0\left[(J(i_{t+1}) - J(i_t))^2\right] \\ &= \frac{1}{2}\sum_{i=1}^{n}\pi(i)\sum_{j=1}^{n}p_{ij}(J(i)-J(j))^2 \\ &= \frac{1}{2}\sum_{i=1}^{n}\pi(i)\sum_{j=1}^{n}p_{ij}(J^2(i)+J^2(j)-2J(i)J(j)) \\ &= \frac{1}{2}\left[\sum_{i=1}^{n}\pi(i)J^2(i)\overbrace{\sum_{j=1}^{n}p_{ij}}^{1} + \sum_{j=1}^{n}J^2(j)\overbrace{\sum_{i=1}^{n}\pi(i)p_{ij}}^{\pi(j)}\right] \end{aligned}$$

$$
\begin{aligned}
& - \left(\sum_{i=1}^{n} \pi(i) J(i) \sum_{j=1}^{n} p_{ij} J(j) \right) \\
= & \frac{1}{2} \left(J'DJ + J'DJ \right) - J'DPJ \\
= & J'DJ - J'DPJ \\
= & J'D(I - P)J
\end{aligned}
$$

Lemma 5.18 *Under Assumption 5.2 and Assumption 5.3, there exists a diagonal matrix $\hat{L}$ with positive diagonal entries, such that the matrix $\hat{L}A$ is negative definite.*

□

Proof of Lemma 5.18

Let $\hat{L}$ be a diagonal matrix with the first diagonal entry equal to some scalar $\hat{l} > 0$ and every other diagonal entry equal to one. Using the special form of the matrix A (see Lemma 5.16) and the negative definiteness of the lower diagonal block of A (see Lemma 5.17) it is a matter of simple algebra to verify that $\hat{L}A$ becomes negative definite for $\hat{l}$ sufficiently large.

□

Note that $\hat{L}A$ negative definite implies A is non-singular. Consider the change of co-ordinates $\tilde{\theta}_t = \hat{L}^{\frac{1}{2}} \theta_t$. We may rewrite the equation 5.19, in terms of $\tilde{\theta}_t$, to obtain

$$
\tilde{\theta}_{t+1} = \tilde{\theta}_t + \gamma_t \left[\hat{L}^{\frac{1}{2}} A(X_t) \hat{L}^{-\frac{1}{2}} \tilde{\theta}_t + \hat{L}^{\frac{1}{2}} b(X_t) \right]
$$

Note that

$$
E_0 \left[\hat{L}^{\frac{1}{2}} A(X_t) \hat{L}^{-\frac{1}{2}} \right] = \hat{L}^{\frac{1}{2}} A \hat{L}^{-\frac{1}{2}} \equiv \hat{A}
$$

$$E_0\left[\hat{L}^{\frac{1}{2}}b(X_t)\right] = \hat{L}^{\frac{1}{2}}b \equiv \hat{b}$$

Note that $\hat{L}A$ negative definite implies $\hat{A}$ is negative definite. Here $\hat{L}^{\frac{1}{2}} = \text{diag}(\hat{l}^{\frac{1}{2}}, 1, \ldots, 1)$ and $\hat{L}^{-\frac{1}{2}} = \text{diag}(\hat{l}^{-\frac{1}{2}}, 1, \ldots, 1)$.

Now we may use the very general result (Theorem 3.2 of Chapter 3) to show that $\tilde{\theta}_t$ converges to the unique solution of

$$\hat{A}\tilde{\theta} + \hat{b} = \underline{0},$$

namely

$$\tilde{\theta}_* = -\hat{A}^{-1}\hat{b}$$

Note that all the assumptions of Theorem 3.2 of Chapter 3 are satisfied (see [54, 55]).

Hence we have θ_t converges to

$$\begin{aligned}
\theta_* &= \hat{L}^{-\frac{1}{2}}\tilde{\theta}_* \\
&= -\hat{L}^{-\frac{1}{2}}\hat{L}^{\frac{1}{2}}A^{-1}\hat{L}^{-\frac{1}{2}}\hat{L}^{\frac{1}{2}}b \\
&= -A^{-1}b,
\end{aligned}$$

i.e. to the unique solution of the linear equation

$$A\theta + b = \underline{0}$$

Hence the following corollary.

Corollary 5.3 *Under Assumption 5.2, Assumption 5.3, Assumption 5.4 and Assumption 5.5, θ_t as defined by equation 5.19 converges to the unique solution of $A\theta + b = \underline{0}$, where A and b are as in Lemma 5.16.* □

Notice that if we use $\tilde{\theta}$ instead of θ in Theorem 3.2 of Chapter 3, we may use the Lyapunov function

$$U(\tilde{\theta}) = \frac{1}{2}(\tilde{\theta} - \tilde{\theta}_*)'(\tilde{\theta} - \tilde{\theta}_*)$$

Also

$$U(\tilde{\theta}) \geq \alpha|\tilde{\theta}|^2, \qquad \text{if} \quad |\tilde{\theta}| > \underbrace{\rho_{0,\alpha}}_{\text{depends on } \alpha}$$

for any $\frac{1}{2} > \alpha > 0$. Here $|\tilde{\theta}|$ denotes the Eucledian norm of $\tilde{\theta}$ as in Chapter 3.

We derive the desired properties of $U(\tilde{\theta})$ next. But before that, we have the following lemma.

Lemma 5.19 *Note that $M \in \mathbf{R}^{n\times n}$ is negative definite implies*

(a) *M^T is negative definite.*

(b) *M^{-1} is negative definite.*

□

Proof of Lemma 5.19

Proof of part (a) is straightforward.

$$x'Mx \;<\; 0, \qquad \forall x \neq \underline{0},\ x \in \mathbf{R}^n,$$

Hence

$$x'M^Tx < 0, \qquad \forall x \neq \underline{0}.$$

Thus M^T is negative definite.

Proof of part (b) is as follows.

$$(M^T)^{-1}M^TM^{-1} = M^{-1}.$$

Hence

$$x'(M^T)^{-1}M^TM^{-1}x = x'M^{-1}x.$$

That is

$$y'M^Ty = x'M^{-1}x \qquad \text{where } y = M^{-1}x.$$

Note that $y = \underline{0}$ if and only if $x = \underline{0}$. Hence by part (a),

$$0 > y'M^Ty = x'M^{-1}x \qquad \text{whenever } x \neq \underline{0}.$$

□

Now back to the properties of $U(\tilde{\theta})$.

$$\begin{aligned} U(\tilde{\theta}) &= \frac{1}{2}(\tilde{\theta} - \tilde{\theta}_*)'(\tilde{\theta} - \tilde{\theta}_*), \\ U'(\tilde{\theta}) &= (\tilde{\theta} - \tilde{\theta}_*), \end{aligned}$$

where $U'(\tilde{\theta})$ is the derivative of $U(\tilde{\theta})$ at $\tilde{\theta}$. Now

$$\begin{aligned} U'(\tilde{\theta}) \cdot (\hat{A}\tilde{\theta} + \hat{b}) &= (\tilde{\theta} + \hat{A}^{-1}\hat{b})'(\hat{A}\tilde{\theta} + \hat{b}) \\ &= (\hat{A}^{-1}(\hat{A}\tilde{\theta} + \hat{b}))'(\hat{A}\tilde{\theta} + \hat{b}) \\ &= (\hat{A}\tilde{\theta} + \hat{b})'(\hat{A}^{-1})^T(\hat{A}\tilde{\theta} + \hat{b}) \\ &< 0, \qquad \text{whenever } \hat{A}\tilde{\theta} + \hat{b} \neq \underline{0}, \end{aligned}$$

that is whenever $\tilde{\theta} \neq \tilde{\theta}_*$. Note that $U(\tilde{\theta}) = 0$ if and only if $\tilde{\theta} = \tilde{\theta}_*$.

Let $0 < \alpha < \frac{1}{2}$. Now

$$U(\tilde{\theta}) - \alpha\tilde{\theta}'\tilde{\theta} = \frac{1}{2}(1 - 2\alpha)\tilde{\theta}'\tilde{\theta} - (\tilde{\theta}_*)'\tilde{\theta} + \frac{1}{2}(\tilde{\theta}_*)'\tilde{\theta}_*.$$

This is a quadratic function with positive definite Hessian matrix and has its minima at $(1-2\alpha)^{-1}\tilde{\theta}_*$ and minimum value

$$\frac{1}{2}|\tilde{\theta}_*|^2 \cdot (1 - \frac{1}{1-2\alpha}).$$

It's Hessian matrix is $(1-2\alpha)I$, where I is the identity matrix. Hence if

$$|\tilde{\theta}| > \overbrace{|\tilde{\theta}_*|(\frac{\sqrt{2\alpha}}{1-2\alpha} + \frac{1}{1-2\alpha})}^{\rho_{0,\alpha}},$$

then $U(\tilde{\theta}) > \alpha\tilde{\theta}'\tilde{\theta}$.

The proof of Theorem 5.1 is exactly similar to that in [55]; that is θ_t converges with probability one to the unique limit θ_* that satisfies $A\theta_* + b = 0$.

Thus $\vartheta_t = \theta_t(1)$ converges with probability one to ϑ^* and r_t converges to $r^* = (\theta_*(2), \ldots, \theta_*(K+1))'$. Note that [55] shows that r^* satisfies

$$\Phi r^* = \Pi\bar{T}^{(\lambda)}(\Phi r^*),$$

that is Φr^* is a fixed point of the operator $\Pi\bar{T}^{(\lambda)}$.We now prove that $\Pi\bar{T}^{(\lambda)}$ has a unique fixed point (which is not proved in [55]).

Lemma 5.20 *$\Pi\bar{T}^{(\lambda)}$ has a unique fixed point.*

□

Proof of Lemma 5.20

Note that Φr^* is a fixed point of $\Pi\bar{T}^{(\lambda)}$ was established in [55]; hence $\Pi\bar{T}^{(\lambda)}$ has a fixed point. To prove uniqueness, first note that any fixed point should be of the form Φr. Let Φr and $\Phi\bar{r}$ be two fixed points, that is

$$\Pi\bar{T}^{(\lambda)}(\Phi r) \quad = \quad \Phi r$$

and

$$\Pi\bar{T}^{(\lambda)}(\Phi\bar{r}) \quad = \quad \Phi\bar{r}.$$

Subtracting the above equations, we get

$$\Pi P^{(\lambda)}\Phi(r-\bar{r}) = \Phi(r-\bar{r}).$$

Hence $P^{(\lambda)}\Phi(r-\bar{r}) - \Phi(r-\bar{r})$ is $D-$orthogonal to ϕ_ks, $k = 1, \ldots, K$. That is

$$\Phi' D\left(P^{(\lambda)}\Phi(r-\bar{r}) - \Phi(r-\bar{r})\right) = \underline{0},$$

the zero vector. Hence

$$\Phi' D(P^{(\lambda)} - I)\Phi(r-\bar{r}) = \underline{0}.$$

But $\Phi' D(P^{(\lambda)} - I)\Phi$ is negative definite (and hence non-singular) by Lemma 5.17. Hence $r - \bar{r} = \underline{0}$ or $r = \bar{r}$. □

Note that $\bar{T}^{(\lambda)}$ has multiple fixed points (Lemma 5.15). It may be shown that $\Pi\bar{T}^{(\lambda)}$ is a contraction mapping under $\|\cdot\|_D$ norm if $\lambda \in (0,1)$ [12, Proposition 6.6.2, pages 381-382] and hence has a unique fixed point. Another proof for the unique fixed point of $\Pi\bar{T}^{(\lambda)}$ when $\lambda = 0$ is given in [12, Proposition 6.6.1, pages 379-381].

5.7.2 Approximation Error

In this sub-section we deal with the approximation error [55, Section 4]. In the context of average cost problem, one is usually content with an approximation of any differential cost J, not necessarily the basic one. We will define the *approximation error* as the infimum of the weighted Eucledian distance from the set of all

differential cost functions.

$$\inf_{J\in\{J^*+c\underline{1}|c\in\mathbf{R}\}} \| \Phi r^* - J \|_D = \inf_{c\in\mathbf{R}} \| \Phi r^* - (J^* + c\underline{1}) \|_D \ .$$

Now any vector $J \in \mathbf{R}^n$ can be decomposed into a component $\boldsymbol{\mathcal{P}}J$ that is $D-$orthogonal to $\underline{1}$, and a component $(I - \boldsymbol{\mathcal{P}})J$ that is a multiple of $\underline{1}$, where $\boldsymbol{\mathcal{P}}$ is the projection matrix defined by

$$\boldsymbol{\mathcal{P}} = I - \underline{1}\underline{1}'D = I - \underline{1}\pi'$$

Note that $\| \underline{1} \|_D = 1$. Also note that for any $J \in \mathbf{R}^n$, $\boldsymbol{\mathcal{P}}J$ is the projection of J under the $\| \cdot \|_D$ metric onto the sub-space which is D-orthogonal to $\underline{1}$.

Also $P\boldsymbol{\mathcal{P}} = \boldsymbol{\mathcal{P}}P$. By the definition of J^*, we have $\pi' J^* = 0$; hence $\boldsymbol{\mathcal{P}}J^* = J^*$. Since for any $r \in \mathbf{R}^K$, $J \in \mathbf{R}^n$, the minimum distance (under the $\| \cdot \|_D$ metric) of the vector $\Phi r - J$ from the sub-space $\{c\underline{1}|c \in \mathbf{R}\}$ is equal to the magnitude of the projection onto the orthogonal complement of the sub-space; we have

$$\inf_{c\in\mathbf{R}} \| \Phi r - (J + c\underline{1}) \|_D = \| \boldsymbol{\mathcal{P}}\Phi r - \boldsymbol{\mathcal{P}}J \|_D$$

In particular

$$\inf_{c\in\mathbf{R}} \| \Phi r^* - (J^* + c\underline{1}) \|_D = \| \boldsymbol{\mathcal{P}}\Phi r^* - J^* \|_D$$

In [55] it is shown that if we replace the basis functions ϕ_k with $\bar{\phi}_k = \boldsymbol{\mathcal{P}}\phi_k$ (which is $D-$orthogonal to $\underline{1}$), the limit to which the TD(λ) converges and the resulting approximation error remains the same. If we let $\bar{\Phi} \equiv \boldsymbol{\mathcal{P}}\Phi$, then $\bar{\Phi}$ also satisfies Assumption 5.3. Letting

$$\bar{\Pi} = \bar{\Phi}(\bar{\Phi}' D \bar{\Phi})^{-1} \bar{\Phi}' D,$$

the projection matrix onto the space spanned by $\bar{\phi}_k$s $\{k = 1, \ldots, K\}$, it may be shown that [55]

$$\bar{\Pi}\bar{T}^{(\lambda)}(\bar{\Phi}r^*) = \bar{\Phi}r^*,$$

where r^* is as before, obtained from the unique solution of $A\theta + b = 0$, where A and b are defined as in Lemma 5.16 (i.e. $r^* = (\theta_*(2), \ldots, \theta_*(K+1))'$, where $A\theta_* + b = 0$). We let

$$\begin{aligned} P_\varsigma &= I + \varsigma(P - I), \\ P_\varsigma^{(\lambda)} &= I + \varsigma(P^{(\lambda)} - I), \end{aligned} \tag{5.22}$$

for $\varsigma \geq 0$, and we define a scalar α_λ for each $\lambda \in [0, 1)$ by

$$\alpha_\lambda \equiv \inf_{\varsigma > 0} \| \bar{\Pi} P_\varsigma^{(\lambda)} \|_D$$

where the norm in the above equation is the induced matrix norm (under the $\| \cdot \|_D$ norm). We have the following error bound [55].

Lemma 5.21 *Let Assumption 5.2 and Assumption 5.3 hold. For each $\lambda \in [0, 1)$, let $r_\lambda^* \in \mathbf{R}^K$ be the unique vector satisfying*

$$\Phi r_\lambda^* = \Pi\bar{T}^{(\lambda)}(\Phi r_\lambda^*).$$

Then

(a) *For each $\lambda \in [0, 1)$, α_λ is in $[0, 1)$ and $\lim_{\lambda \uparrow 1} \alpha_\lambda = 0$.*

(b) *The following bound holds:*

$$\| \boldsymbol{\mathcal{P}}\Phi r_\lambda^* - J^* \|_D \leq \frac{1}{\sqrt{1 - \alpha_\lambda^2}} \inf_{r \in \mathbf{R}^K} \| \boldsymbol{\mathcal{P}}\Phi r - J^* \|_D .$$

□

Note that the bound is a multiple of

$$\inf_{r\in\mathbf{R}^K} \| \boldsymbol{\mathcal{P}}\Phi r - J^* \|_D,$$

which is the minimal error possible, given the fixed set of basis functions. This term becomes zero if there exists a parameter vector r and a scalar c for which $\Phi r = J^* + c\underline{1}$, that is, if the "approximation architecture" is capable of representing exactly some differential cost function. Note that because of Assumption 5.3(b) this r is unique, if at all such an r exists. That is if $\{J^* + c\underline{1} | c \in \mathbf{R}\}$ intersects the space spanned by ϕ_ks for $k = 1, 2, \ldots, K$; then it intersects at a unique point. Note that in this case, by Lemma 5.15 and the definition of $\bar{\Pi}$, $r = r^*$.

The proof of Lemma 5.21 is exactly as in the proof of [55, Theorem 3]; but we need to prove $\lim_{\lambda\uparrow 1} \alpha_\lambda = 0$, for the case when the transition probability matrix P corresponds to a general (not necessarily aperiodic) Markov chain. As in [55]

$$\begin{aligned}
\limsup_{\lambda\uparrow 1} \alpha_\lambda &= \limsup_{\lambda\uparrow 1} \inf_{\varsigma>0} \| \bar{\Pi} P_\varsigma^{(\lambda)} \|_D \\
&\leq \limsup_{\lambda\uparrow 1} \| \bar{\Pi} P^{(\lambda)} \|_D \leq \limsup_{\lambda\uparrow 1} \| \boldsymbol{\mathcal{P}} P^{(\lambda)} \|_D,
\end{aligned}$$

the last inequality follows from the fact that $\bar{\Pi}$ projects onto a subspace of $\underline{1}_\perp$ (the sub-space onto which $\boldsymbol{\mathcal{P}}$ projects), that is $\bar{\Pi} = \bar{\Pi}\boldsymbol{\mathcal{P}}$ and that projection does not increase the norm.

Since $P^* = \underline{1}\pi'$, we have $\boldsymbol{\mathcal{P}}P^* = \mathbf{0}$, the zero matrix. Based on the discussion following Lemma 5.14, for $\lambda \in (0, 1)$,

$$P^{(\lambda)} = \frac{1-\lambda}{\lambda} \left((I - \lambda P)^{-1} - I \right)$$

Also by Lemma 5.1,

$$(I - \lambda P)^{-1} = (1 - \lambda)^{-1} P^* + L + \mathcal{O}(|1 - \lambda|),$$

where P^* is the limiting matrix corresponding to P and L is the differential matrix corresponding to P, given by

$$L = (I - P + P^*)^{-1} - P^*$$

and $\lim_{\lambda \to 1} \mathcal{O}(|1 - \lambda|) = \mathbf{0}$. Hence

$$\begin{aligned} \boldsymbol{\mathcal{P}} P^{(\lambda)} &= \frac{1 - \lambda}{\lambda} \boldsymbol{\mathcal{P}} \left[(1 - \lambda)^{-1} P^* + L + \mathcal{O}(|1 - \lambda|) - I \right] \\ &= \frac{1 - \lambda}{\lambda} \boldsymbol{\mathcal{P}} (L - I) + \boldsymbol{\mathcal{P}} \mathcal{O}(|1 - \lambda|) \end{aligned}$$

which tends to $\mathbf{0}$ as $\lambda \uparrow 1$. Hence

$$\lim_{\lambda \uparrow 1} \| \boldsymbol{\mathcal{P}} P^{(\lambda)} \|_D = 0.$$

Thus $\lim_{\lambda \uparrow 1} \alpha_\lambda = 0$.

5.7.3 Using A Fixed Average Cost Estimate

In this subsection, we introduce, as in [55], a variant of the temporal difference scheme that employs a fixed estimate ϑ of the average cost, in place of ϑ_t. In particular the parameter vector r_t is updated according to the same rule 5.16, but the definition of the temporal difference equation 5.14 is changed to

$$d_t = (g_t - \vartheta) + \phi'(i_{t+1}) r_t - \phi'(i_t) r_t$$

Also define

$$\tilde{\alpha}_\lambda \equiv \inf_{\varsigma \in [0,1]} \| \Pi P_\varsigma^{(\lambda)} \|_D,$$

where the norm in the above equation is the induced matrix norm (under the $\|\cdot\|_D$ norm). Then we have the following lemma [55, Theorem 4]

Lemma 5.22 *Under Assumptions 5.2–5.5, for any $\lambda \in [0,1)$, the following hold:*

(a) *The* TD*(λ) algorithm with a fixed average cost estimate, as defined above converges with probability one.*

(b) *The limit of convergence $\bar{r}_\lambda$ is the unique solution of the equation*

$$\Pi \bar{T}^{(\lambda)}(\Phi \bar{r}_\lambda) + \frac{\vartheta^* - \vartheta}{1-\lambda} \Pi \underline{1} = \Phi \bar{r}_\lambda.$$

(c) *For any $\lambda \in [0,1)$, $\tilde{\alpha}_\lambda$ is in $[\| \underline{1} \|_D, 1)$, and*

$$\lim_{\lambda \uparrow 1} \tilde{\alpha}_\lambda = \| \Pi \underline{1} \|_D \ .$$

(d) *The limit of convergence $\bar{r}_\lambda$ satisfies*

$$\begin{aligned} \| \boldsymbol{\mathcal{P}} \Phi \bar{r}_\lambda - J^* \|_D \ &\leq \ \frac{1}{\sqrt{1-\alpha_\lambda^2}} \inf_{r \in \mathbf{R}^K} \| \boldsymbol{\mathcal{P}} \Phi r - J^* \|_D \\ &\quad + \frac{|\vartheta^* - \vartheta|}{(1-\tilde{\alpha}_\lambda)(1-\lambda)} \| \Pi \underline{1} \|_D, \end{aligned}$$

where α_λ and $\boldsymbol{\mathcal{P}}$ are defined as earlier.

□

Proof of Lemma 5.22

Note that we are dealing with the general (aperiodic or periodic) Markov chain. We omit the proofs of parts (a)–(b), because it is very similar to the proof of Theorem 5.1.

Note that there is some correction to part (c) from [55]. The proof that $\tilde{\alpha}_\lambda < 1$, is similar to that of Lemma 5.21, part (a) [55, Theorem 3]. However it needs a bit more explanation. From Lemma 5.14 we have $\| P^{(\lambda)} \|_D \leq 1$, and $P^{(\lambda)} J \neq J$ if J is not proportional to $\underline{1}$. It follows that for any $\varsigma \in (0,1)$ and J that is not proportional to $\underline{1}$, we have

$$\| \Pi P_\varsigma^{(\lambda)} J \|_D \leq \| P_\varsigma^{(\lambda)} J \|_D = \| \varsigma P^{(\lambda)} J + (1-\varsigma) J \|_D < \| J \|_D \, .$$

The first inequality uses the non-expansive property of the projection. The last one holds because J and $P^{(\lambda)} J$ are distinct elements of the ball $\{\bar{J} | \, \| \bar{J} \|_D \leq \| J \|_D\}$, so their strict convex combination must lie in the interior interior.

Also note that $\| \Pi \underline{1} \|_D < 1$, since $\| \underline{1} \|_D = 1$, and $\Pi \underline{1} + (\underline{1} - \Pi \underline{1}) = \underline{1}$. Note that by Assumption 5.3(b), $(\underline{1} - \Pi \underline{1}) \neq \underline{0}$; also $\Pi \underline{1}$ is D-orthogonal to $(\underline{1} - \Pi \underline{1})$. Hence by Pythagorean theorem

$$1 = \| \underline{1} \|_D^2 = \| \Pi \underline{1} \|_D^2 + \| \underline{1} - \Pi \underline{1} \|_D^2,$$

and hence $\| \Pi \underline{1} \|_D < 1$. Hence if $J = c\underline{1}$, with $|c| \leq 1$, then $\| \Pi J \|_D < 1$.

(Note that $\| \Pi P_\varsigma^{(\lambda)} J \|_D$ is a continuous function of J and that the set $\{J | \, \| J \|_D \leq 1\}$ is compact. Thus for any $\varsigma \in (0,1)$, $\| \Pi P_\varsigma^{(\lambda)} \|_D < 1$. Since

$$\tilde{\alpha}_\lambda = \inf_{\varsigma \in [0,1]} \| \Pi P_\varsigma^{(\lambda)} \|_D \leq \inf_{\varsigma \in (0,1)} \| \Pi P_\varsigma^{(\lambda)} \|_D < 1.$$

Also for any ς, $\Pi P_\varsigma^{(\lambda)} \underline{1} = \Pi \underline{1}$. Thus $\tilde{\alpha}_\lambda \geq \| \Pi \underline{1} \|_D$. Now

$$\begin{aligned} \limsup_{\lambda \uparrow 1} \tilde{\alpha}_\lambda &= \limsup_{\lambda \uparrow 1} \inf_{\varsigma \in [0,1]} \| \Pi P_\varsigma^{(\lambda)} \|_D \\ &\leq \limsup_{\lambda \uparrow 1} \| \Pi P^{(\lambda)} \|_D \end{aligned}$$

Based on Lemma 5.1 and the discussion following Lemma 5.14, for any $\lambda \in (0,1)$,

$$\begin{aligned}\Pi P^{(\lambda)} &= \frac{1-\lambda}{\lambda}\Pi\left[(I-\lambda P)^{-1} - I\right] \\ &= \frac{1-\lambda}{\lambda}\Pi\left[(1-\lambda)^{-1}P^* + L + \mathcal{O}(|1-\lambda|) - I\right].\end{aligned}$$

Here P^* and L are the limiting matrix and differential matrix, respectivelty of the stochastic matrix P. Hence

$$\lim_{\lambda\uparrow 1}\Pi P^{(\lambda)} = \Pi P^*.$$

Now $P^* = \underline{1}\pi'$.

$$\begin{aligned}\|\,\Pi P^*\,\|_D &= \sup_{\|J\|_D=1}\|\,\Pi\underline{1}\pi' J\,\|_D \\ &= \|\,\Pi\underline{1}\,\|_D \sup_{\|J\|_D=1}|\pi' J|\end{aligned}$$

Now any J can be decomposed as

$$J = \boldsymbol{\mathcal{P}}J + (\pi' J)\underline{1}$$

the two terms being D-orthogonal. By Pythagorean theorem,

$$\|\,J\,\|_D^2 = \|\,\boldsymbol{\mathcal{P}}J\,\|_D^2 + |\pi' J|^2\,\|\,\underline{1}\,\|_D^2\,.$$

Hence if $\|\,J\,\|_D = 1$, then $|\pi' J| \le 1$ (note $\|\,\underline{1}\,\|_D = 1$). If $J = \underline{1}$, $|\pi' J| = 1$. Hence

$$\|\,\Pi P^*\,\|_D = \|\,\Pi\underline{1}\,\|_D$$

Thus

$$\limsup_{\lambda\uparrow 1}\tilde{\alpha}_\lambda \le \limsup_{\lambda\uparrow 1}\|\,\Pi P^{(\lambda)}\,\|_D = \|\,\Pi\underline{1}\,\|_D\,.$$

Hence

$$\lim_{\lambda\uparrow 1}\tilde{\alpha}_\lambda = \|\,\Pi\underline{1}\,\|_D$$

Note that $\Pi \underline{1} = \underline{0}$ if and only if $\Pi = \bar{\Pi}$. That is the basis functions ϕ_ks (for $k = 1, 2, \ldots, K$) lie in $\underline{1}_\perp$ (the sub-space which is D-orthogonal to $\underline{1}$), since $(\underline{1} - \Pi \underline{1})$ is D-orthogonal to ϕ_ks (for $k = 1, 2, \ldots, K$).

Proof of part (d) of the Lemma 5.22 is as in [55].

□

Note that $\tilde{\alpha}_\lambda$ could have been defined as $\inf_{\varsigma > 0} \| \Pi P^{(\lambda)} \|_D$, and all the results of Lemma 5.22 hold.

Also note that for the TD(λ) scheme, the initial eligibility vector z_{-1}, could be any value, not necessarily $\underline{0}$. Note that η_t could be any non-negative deterministic sequence satisfying

$$\sum_{t=0}^{\infty} \eta_t = \infty; \ \sum_{t=0}^{\infty} \eta_t^2 < \infty, \ \text{and} \ \sum_{t=0}^{\infty} |\eta_{t+1} - \eta_t| < \infty,$$

and the estimate of the average cost ϑ_t will converge with probability one to ϑ^*. If η_t is a non-increasing sequence then $\sum_{t=0}^{\infty} |\eta_{t+1} - \eta_t| < \infty$ is satisfied.

In the TD(λ) algorithm, η_t need not be $c\gamma_t$, but any deterministic sequence which satisfies the above property, and Theorem 5.1 holds. Notice how ϑ_t enters the computation of r_t, from equation 5.19. We are not providing the rigorous proof of this, but may be inferred from Lemma 5.22 (a) and (b). Also the moment condition on immediate costs in Assumption 5.5, namely

$$E\left[|g_t|^k | i_t = i\right] < \infty,$$

$\forall i \in \mathcal{S}$, need be satisfied only upto a sufficiently large k ($k = 4$) and not for all $k > 0$. (see Chapter 3, Proposition 3.1).

We have the following useful lemma which comes in handy later.

Lemma 5.23 *Let $\mathcal{R}$ be an $n-1$ dimensional subspace of $\mathbf{R}^n$, such that $\underline{1} \notin \mathcal{R}$. Let J be an arbitrary vector in $\mathbf{R}^n$. Then the line $\{J + c\underline{1} | c \in \mathbf{R}\}$, intersects the subspace $\mathcal{R}$ at a unique point.*

□

Proof of Lemma 5.23

Let $v_1, v_2, \ldots, v_{n-1}$ be a basis for $\mathcal{R}$; that is, the v_is are non-zero vectors which are linearly independent and span $\mathcal{R}$. Since $v_1, v_2, \ldots, v_{n-1}$ together with $\underline{1}$ form a basis for $\mathbf{R}^n$,

$$J = \sum_{l=1}^{n-1} c_l v_l + c_n \underline{1},$$

where c_ls are scalar values in $\mathbf{R}$. Note that the c_ls are unique. Let

$$\bar{J} = \sum_{l=1}^{n-1} c_l v_l.$$

$\bar{J} \in \mathcal{R}$ and $\bar{J} = J - c_n \underline{1}$. Hence $\bar{J} \in \mathcal{R} \cap \{J + c\underline{1} | c \in \mathbf{R}\}$.

To prove the uniqueness of $\bar{J}$, suppose $\hat{J} \in \mathcal{R} \cap \{J + c\underline{1} | c \in \mathbf{R}\}$ and $\hat{J} \neq \bar{J}$. Hence $\hat{J} - \bar{J}$ lies in $\mathcal{R}$, and is a non-zero scalar multiple of $\underline{1}$, which leads to a contradiction, since we assume that $\underline{1}$ does not lie in the subspace $\mathcal{R}$. Note that $\bar{J}$ does not depend on the choice of the basis for $\mathcal{R}$. For example let $w_1, w_2, \ldots, w_{n-1}$ be another basis for $\mathcal{R}$. Let $\hat{c}_1, \ldots, \hat{c}_{n-1}$ be such that, $\sum_{l=1}^{n-1} \hat{c}_l w_l = \bar{J}$, which is possible since $\bar{J} \in \mathcal{R}$. Note that

$$J = \bar{J} + \hat{c}_n \underline{1} = \sum_{l=1}^{n-1} \hat{c}_l w_l + \hat{c}_n \underline{1}.$$

where $\hat{c}_n = c_n$. Note that the $\hat{c}_l$ are unique, and

$$\bar{J} = J - \hat{c}_n \underline{1} \in \mathcal{R} \cap \{J + c\underline{1} | c \in \mathbf{R}\}.$$

□

Corollary 5.4 *In particular if in Assumption 5.3, $K = n-1$; that is $\phi_1, \phi_2, \cdots, \phi_{n-1} \in \mathbf{R}^n$ are linearly independent and $\underline{1}$ is not in the span of $\phi_1, \ldots, \phi_{n-1}$; then the line $\{J^* + c\underline{1} | c \in \mathbf{R}\}$ intersects the subspace $\{\Phi r | r \in \mathbf{R}^K\}$ at a unique point $\bar{J}$, which is the unique fixed point of the operator $\Pi \bar{T}^{(\lambda)}$ (see Lemma 5.15). Hence $\bar{J} = \Phi r^*$, where r^* is as in Theorem 5.1. Note that Φ is an $n \times K$ matrix where the k^{th} column is ϕ_k.*

□

Corollary 5.5 *Let $e_k = (0, \ldots, 0, \overbrace{1}^{k^{\text{th}} \text{ entry}}, 0, \ldots, 0)'$, the k^{th} standard basis vector in $\mathbf{R}^n$. Fix $m \in \{1, \ldots, n\}$. Let $\phi_l = e_{i_l}$, for $l = 1, \ldots, n-1$, where $i_l \in \{1, \ldots, n\} \backslash \{m\}$ (that is $i_l \in \{1, \ldots, n\}$, but not m). Here $i_l \neq i_{\tilde{l}}$, if $l \neq \tilde{l}$.*

Then $\{J^* + c\underline{1} | c \in \mathbf{R}\}$ intersects the span of $\{\phi_1, \ldots, \phi_{n-1}\}$ at a unique point $\bar{J}$, which is the unique fixed point of operator $\Pi \bar{T}^{(\lambda)}$. In particular $e_i' \Phi r^* = \bar{J}(i) = J^*(i) - J^*(m)$, for $i \in \{1, \ldots, n\}$.

□

5.8 Stationary Randomized Policies

Please see the Section 4.2 in Chapter 3.

Suppose all the stationary deterministic policies are unichain; then any stationary randomized policy is unichain and the recurrent class for a stationary fully randomized policy is the union of the recurrent classes of all the stationary deterministic policies (see Appendix B). Hence in this case, the limiting matrix P_δ^* and differential matrix L_δ corresponding to P_δ, where

$$\begin{aligned} P_\delta^* &= \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} P_\delta^k, \\ L_\delta &= (I - P_\delta + P_\delta^*)^{-1} - P_\delta^*, \end{aligned}$$

are continuous functions on the space Λ of stationary randomized policies (see Lemma 5.10). In fact compactness of Λ implies that P_δ^* and L_δ are uniformly continuous on Λ for a unichain MDP. In particular given any $\epsilon > 0$, there exists a $\varepsilon > 0$ (dependent on ϵ) such that $\| P_\delta^* - P_\mu^* \|_\infty < \epsilon$, $\| L_\delta - L_\mu \|_\infty < \epsilon$ for each $\mu \in \Upsilon$ and $\delta \in \Lambda$, with $\mathsf{d}(\mu, \delta) < \varepsilon$, where $\| \cdot \|_\infty$ is the vector norm on matrices, defined by equation 5.11 in Section 5.5, and d is the metric defined on Λ (see the section on Stationary Randomized Policies in Chapter 4).

5.9 TD For Learning

Here we are interested in learning the optimal average cost, an optimal policy (stationary deterministic) and associated differential cost function for a unichain MDP with a common recurrent state (see Lemma 5.4). Neither the transition probabilities nor the distribution of (or expected value of) immediate costs are known in advance.

In the following, $\mathcal{Q}$ represents the set of feasible state-action pairs.

5.9.1 Recurrent MDPs

We first give an algorithm for recurrent MDPs.

Assumption 5.6 *Let the following hold.*

(a) *The monotonic step sizes $\gamma_t \downarrow 0$, are pre-determined non-negative scalars and satisfy $\sum_{t=0}^{\infty} \gamma_t = \infty$; $\sum_{t=0}^{\infty} \gamma_t^2 < \infty$. Let $\eta_t = c\gamma_t$ for some positive real value c.*

(b) *Let $K = n - 1$, and without loss of generality*

$$\phi_k = (0, \ldots, 0, \overbrace{1}^{k^{\text{th}} \text{ entry}}, 0, \ldots, 0)',$$

the k^{th} standard basis vector in $\mathbf{R}^n$; implying lookup table representation.

(c) *The immediate cost g_t, has finite moments, that is*

$$E\left[|g_t|^k | i_t = i, u_t = a\right] < \infty,$$

$\forall i \in \mathcal{S}, a \in \mathcal{A}(i), \forall k \in \mathbf{N}$.

(d) *For each state action pair $(i, a) \in \mathcal{Q}$, let the pre-determined scalar non-negative step sizes $\gamma_t(i, a)$ be such that*

$$\sum_{t=1}^{\infty} \gamma_t(i, a) = \infty, \quad \sum_{t=1}^{\infty} (\gamma_t(i, a))^2 < \infty.$$

□

Assumption 5.7 *The MDP is recurrent, that is under each stationary deterministic policy, the corresponding Markov Chain is irreducible (it may be aperiodic or periodic).* □

Assumption 5.7 says that any $\delta \in \Lambda$, gives rise to an irreducible Markov chain with unique invariant distribution π^δ where $\pi^\delta(i) > 0,\ \forall i \in \mathcal{S}$. Fix a policy $\delta \in \Lambda$. We want to estimate the average cost $\vartheta^\delta = (\pi^\delta)'\bar{g}^\delta$ and a differential cost J_δ, which is the unique point in $\{J_\delta^* + \tilde{c}\underline{1} | \tilde{c} \in \mathbf{R}\} \cap \{\Phi r | r \in \mathbf{R}^K\}$, given by

$$J_\delta(i) \quad = \quad J_\delta^*(i) - J_\delta^*(n), \quad i \in \mathcal{S} \tag{5.23}$$

Here J_δ^* is the basic differential cost for policy δ. We use $\tilde{J}(\cdot, r) = \Phi r$, to approximate J_δ. Note that for our choice of ϕ_ks, $\tilde{J}(i, r) = r'\phi(i) = r(i),\ i \in \{1, \ldots, n-1\}$ and $\tilde{J}(n, r) = 0$.

Note that instead of state n, we could have chosen any state $\hat{n} \in \mathcal{S}$, with an appropriate choice of ϕ_ks (leaving out $e_{\hat{n}}$ instead of e_n to form the set of $n-1$ basis functions).

In the following three algorithms in this subsection, we assume that Assumption 5.7 holds.

Fix a policy $\delta \in \Lambda$. We want to estimate the average cost $\vartheta^\delta \in \mathbf{R}$ and a differential cost $J^\delta \in \mathbf{R}^n$ corresponding to stationary policy δ. Let i_t and a_t be the state and action taken at time $t \in \mathbf{N}_0$, while using polcy δ. Let g_t be the corresponding immediate cost incurred.

Algorithm 5.1 *The update rule is as follows (starting at $t = 0$):*

$$\begin{aligned} d_t &= g_t - \vartheta_t + \tilde{J}(i_{t+1}, r_t) - \tilde{J}(i_t, r_t) \\ z_t &= \lambda z_{t-1} + \phi(i_t) \\ \vartheta_{t+1} &= (1 - \eta_t)\vartheta_t + \eta_t g_t \end{aligned}$$

$$
\begin{aligned}
r_{t+1} &= r_t + \gamma_t d_t z_t \\
t &= t+1
\end{aligned}
$$

□

Under Assumption 5.6(a),(b),(c) and Assumption 5.7, in Algorithm 5.1, $\tilde{J}(\cdot, r_t) \rightarrow J_\delta$ almost surely (where J_δ is given by equation 5.23) by Theorem 5.1 and Corollary 5.5. Similarly $\vartheta_t \rightarrow \vartheta^\delta$. Note that the convergence holds irrespective of the initialization r_0, z_{-1}, ϑ_0 or initial state i_0.

In the following algorithm, we try to estimate by simulation, the Q-values corresponding to one step look ahead, with terminal cost $J \in \mathbf{R}^n$, that is

$$Q^J(i,a) = g(i,a) + \sum_{j=1}^{n} p_{ij}(a) J(j)$$

(With slight abuse of notation for the Q-values, we use J instead of a policy $\delta \in \Lambda$, as the superscript).

Fix a stationary fully randomized policy $\delta \in \text{Interior}(\Lambda)$ and a vector $J \in \mathbf{R}^n$ (refer the section on Stationary Randomized Policies in Chapter 4, for the definition of Interior(Λ)). Let i_t and a_t be the state and action taken at time $t \in \mathbf{N}_0$, while using policy δ. Let g_t be the corresponding immediate cost incurred.

Algorithm 5.2

$$\tau_{-1}(i,a) = 0, \quad \forall (i,a) \in \mathcal{Q}$$

The update rule is as follows (starting at $t = 0$):

$$\tau_t(i_t, a_t) = \tau_{t-1}(i_t, a_t) + 1$$

$$
\begin{aligned}
\tau_t(i,a) &= \tau_{t-1}(i,a), \qquad \forall (i,a) \neq (i_t,a_t), (i,a) \in \mathcal{Q} \\
Q_{t+1}(i_t,a_t) &= Q_t(i_t,a_t) + \gamma_{\tau_t(i_t,a_t)} \left(g_t + J(i_{t+1}) - Q_t(i_t,a_t) \right) \\
Q_{t+1}(i,a) &= Q_t(i,a), \qquad \forall (i,a) \neq (i_t,a_t), (i,a) \in \mathcal{Q} \\
t &= t+1
\end{aligned}
$$

□

Here $\tau_t(i,a)$ represents the number of times, action a has been taken from state i, by time $t \in \mathbf{N}_0$. Since policy δ is fully randomized, each state-action pair in $\mathcal{Q}$ is visited infinitely often, as $t \to \infty$. Hence standard results from stochastic approximation theory [16] can be used to show that $Q_t \to Q^J$ almost surely. The convergence holds irrespective of the initial value Q_0 or initial state i_0. All that is required of the non-negative step size parameters $\gamma_t(i,a)$ (for each $(i,a) \in \mathcal{Q}$) is that they should satisfy the standard assumptions

$$
\sum_{t=1}^{\infty} \gamma_t(i,a) = \infty; \quad \sum_{t=1}^{\infty} ((\gamma_t(i,a))^2 < \infty,
$$

almost surely and may be allowed to be random and can depend on past history (at the time the step size is used). Note that the constraint on the moments of g_t, for Algorithm 5.2 is that

$$
E\left[|g_t|^2 | i_t = i, a_t = a \right] < \infty \quad \forall (i,a) \in \mathcal{Q}.
$$

Consider the following algorithm. Here the inequality applied to vectors is componentwise.

Algorithm 5.3

Let $\underline{0} < \bar{\epsilon}_k < \tilde{\epsilon}$ *be a sequence of positive vectors in* $\mathbf{R}^n$*; where* $\tilde{\epsilon} \in \mathbf{R}^n$ *is defined as a vector with* $\tilde{\epsilon}(i) = \frac{1}{|\mathcal{A}(i)|}$.

1. *Set* $k = 0$.

2. *Select an arbitrary stationary deterministic policy* $\mu_0 \in \Upsilon$.

3. *With policy* μ_k*, run Algorithm 5.1, for "large" random number* n_k *of steps, till* Φr_{n_k} *"nearly" converges to* J_{μ_k}*. Let* $J_k = \Phi r_{n_k}$.

4. *Choose the stationary fully randomized extremal policy* $\delta_k \in \Lambda_{\bar{\epsilon}_k}$ *, associated with* μ_k *and run Algorithm 5.2 with one step terminal cost* J_k*, for "large" random number* $\tilde{n}_k$ *of steps, till* $Q_{\tilde{n}_k}$ *"nearly" converges to* Q^{J_k}*. Let* $\tilde{Q} = Q_{\tilde{n}_k}$*. Let*

$$\zeta_k = \max\{\max_{i \in \mathcal{S}} |J_k(i) - J_{\mu_k}(i)|, \max_{(i,a) \in \mathcal{Q}} |\tilde{Q}(i,a) - Q^{J_k}(i,a)|\}.$$

5. *Set* $k = k + 1$ *and update the policy to* μ_k*, where*

$$\mu_k(i) = \arg \min_{a \in \mathcal{A}(i)} \tilde{Q}(i,a)$$

6. *Go to step 3.*

□

We note that the initial condition r_0, z_{-1} and ϑ_0, when calling Algorithm 5.1 in step 3 may be arbitray, but can be set to the final values obtained in the previous iteration, if needed. Similarly i_0, when calling Algorithm 5.1 in step 3 may also be

arbitrary, but may be set to the final value of state obtained in step 4 in the previous iteration.

Similarly the initial condition Q_0, when calling Algorithm 5.2 in step 4 may be arbitrary, but can be set to the final value obtained in the previous iteration if needed. When calling Algorithm 5.2 in step 4, i_0 may be arbitrary, but can be set to the final value of state obtained in step 3 of the current iteration.

In the following theorem and the next one, the following notations hold:

$J^\diamond$ is defined as the unique vector satisfying

$$\vartheta^* \underline{1} + J^\diamond = \bar{T} J^\diamond, \quad J^\diamond(n) = 0.$$

Here ϑ^* is the optimal average cost.

Also for any $\delta \in \Lambda$, J_δ is the unique vector satisfying

$$\vartheta^\delta \underline{1} + J_\delta = \bar{T}_\delta J_\delta, \quad J_\delta(n) = 0.$$

Here ϑ^δ is the average cost corresponding to δ. Note that $J_\delta = J_\delta^* - J_\delta^*(n)\underline{1}$, where J_δ^* is the basic differential cost corresponding to δ. In the following, $\|\cdot\|$ denotes the Eucledian norm for a vector in $\mathbf{R}^n$.

Theorem 5.2 *Consider the Algorithm 5.3, and let Assumption 5.6 and Assumption 5.7 hold. Then we have the following results.*

1. *Given any scalar $\varepsilon > 0$, there exists scalar $\zeta > 0$ and vector $\bar{\epsilon}$, with $\underline{0} < \bar{\epsilon} < \tilde{\epsilon}$, such that if $\bar{\epsilon}_k < \bar{\epsilon}$ and $\zeta_k < \zeta$, $\forall k$; then μ_k "converges" to an optimal stationary deterministic policy (if there are multiple stationary deterministic policies that are optimal, it may take any of them) in a finite number of steps.*

In particular J_{μ_k} converges to $J^\diamond$ and ϑ^{μ_k} converges to ϑ^* in a finite number of steps ($\leq |\Upsilon|$). Also

$$|\vartheta^{\delta_k} - \vartheta^{\mu_k}| < \varepsilon, \qquad \textit{and} \quad \max_{i \in \mathcal{S}} (J_{\delta_k}(i) - J_{\mu_k}(i)) < \varepsilon, \quad \forall k.$$

2. *In particular if* $\limsup_{k\to\infty} \bar{\epsilon}_k(i) = 0, \ \forall i \in \mathcal{S}$, *and* $\limsup_{k\to\infty} \zeta_k = 0$, *then* $\| J_{\mu_k} - J^\diamond \| \to 0$, $|\vartheta^{\mu_k} - \vartheta^*| \to 0$, $\| J_{\delta_k} - J^\diamond \| \to 0$ *and* $\vartheta^{\delta_k} \to \vartheta^*$. *Infact* $J_{\mu_k} = J^\diamond$ *and* $\vartheta^{\mu_k} = \vartheta^*$, *for all large* k.

□

Proof of Theorem 5.2

Proof of part (1) follows from Lemma 5.4, Proposition 5.2, Lemma 5.10 and the arguments following Corollary 5.2. See also Lemma 5.8 and Lemma 5.9.

Proof of part (2) follows from part (1) and arguments along the same line as in the proof of part (1).

□

Note that instead of using an extremal policy $\delta_k \in \Lambda_{\bar{\epsilon}_k}$ to approximate μ_k, in Algorithm 5.3, we could have chosen $\tilde{\delta}_k \in \text{Interior}(\Lambda)$ such that

$$[\tilde{\delta}_k(i)]_{\mu_k(i)} \geq (1 - (|\mathcal{A}(i)| - 1)\bar{\epsilon}_k(i)), \quad \forall i \in \mathcal{S};$$

for instance $\tilde{\delta}_k$ could be made to depend on the approximation to the Q-values obtained in the previous iteration. Also the step size parameters used in step 3 and step 4 of Algorithm 5.3 could vary for different policy evaluations and Q-value computations (i.e. different iterations).

Note that we could have used as basis vectors (ϕ_ks), any $n-1$ vectors whose span does not contain $\underline{1}$, instead of e_ks. However in this case, in Algorithm 5.3, J_k approximates the differential cost J_{μ_k} which is the unique element of $\left\{J^*_{\mu_k} + \tilde{c}\underline{1} | \tilde{c} \in \mathbf{R}\right\} \cap \{\Phi r | r \in \mathbf{R}^K\}$. Here $J^*_{\mu_k}$ is the basic differential cost corresponding to policy μ_k. With a similar definition of J_{δ_k}, J_{δ_k} and J_{μ_k} converges to the unique point in $\{J^\diamond + \tilde{c}\underline{1} | \tilde{c} \in \mathbf{R}\} \cap \{\Phi r | r \in \mathbf{R}^K\}$.

5.9.2 Communicating Unichain MDP With A Common Recurrent State

Assumption 5.8 *The MDP is unichain, has a common recurrent state and is communicating.* □

The communicating property in the above assumption is equivalent to the fact that for any stationary fully randomized policy δ ($\in$ Interior(Λ)), the corresponding Markov chain is irreducible. Any stationary fully randomized policy gives rise to an irreducible Markov chain with the same period and the same periodic classes. Note that the periods of the recurrent classes of the Markov chains corresponding to the various stationary deterministic policies are irrelevant.

In the following algorithms in this subsection, we assume that Assumption 5.6 and Assumption 5.8 hold.

We fix a stationary fully randomized policy $\delta \in$ Interior(Λ). We want to estimate the average cost $\vartheta^\delta = (\pi^\delta)' \bar{g}^\delta$, and a differential cost J_δ which is the unique

point in $\{J_\delta^* + \tilde{c}\underline{1} | \tilde{c} \in \mathbf{R}\} \cap \{\Phi r | r \in \mathbf{R}^K\}$, given by $J_\delta(i) = J_\delta^*(i) - J_\delta^*(n)$, where J_δ^* is the basic differential cost corresponding to δ. We use $\tilde{J}(\cdot, r) = \Phi r$ to approximate $J_\delta(\cdot)$. Note that with our choice of ϕ_ks, $\tilde{J}(i, r) = r'\phi(i) = r(i)$ for $i \in \{1, 2, \ldots, n-1\}$ and $\tilde{J}(n, r) = 0$.

Note that instead of state n, we could have chosen any state $\hat{n} \in \mathcal{S}$, with an appropriate choice of ϕ_ks (leaving out $e_{\hat{n}}$ instead of e_n to form the set of $n-1$ basis functions).

We also want to estimate the Q-values for the policy δ, given by

$$Q^\delta(i, a) = g(i, a) + \sum_{j=1}^{n} p_{ij}(a) J_\delta(j), \quad \forall (i, a) \in \mathcal{Q}.$$

Algorithm 5.4

$$\tau_{-1}(i, a) = 0, \quad \forall (i, a) \in \mathcal{Q}$$

The update rule is as follows (starting at $t = 0$):

$$\begin{aligned}
d_t &= g_t - \vartheta_t + \tilde{J}(i_{t+1}, r_t) - \tilde{J}(i_t, r_t) \\
z_t &= \lambda z_{t-1} + \phi(i_t) \\
\vartheta_{t+1} &= (1 - \eta_t)\vartheta_t + \eta_t g_t \\
r_{t+1} &= r_t + \gamma_t d_t z_t \\
\tau_t(i_t, a_t) &= \tau_{t-1}(i_t, a_t) + 1 \\
\tau_t(i, a) &= \tau_{t-1}(i, a), \qquad \forall (i, a) \neq (i_t, a_t), (i, a) \in \mathcal{Q} \\
Q_{t+1}(i_t, a_t) &= Q_t(i_t, a_t) + \gamma_{\tau_t(i_t, a_t)} \Big(g_t + \tilde{J}(i_{t+1}, r_t) - Q_t(i_t, a_t) \Big) \\
Q_{t+1}(i, a) &= Q_t(i, a), \qquad \forall (i, a) \neq (i_t, a_t), (i, a) \in \mathcal{Q}
\end{aligned}$$

□

Note that z_{-1} need not be the zero vector but may take any arbitrary value. The initial values Q_0, ϑ_0 and r_0 can be arbitrary. Here $\tau_t(i,a)$ represents the number of times action a has been taken from state i by the time $t \in \mathbf{N}_0$. Under Assumption 5.6 and Assumption 5.8, $\tilde{J}(\cdot, r_t) \to J_\delta$, almost surely by Theorem 5.1 and Corollary 5.5. In addition since all the state-action pairs in $\mathcal{Q}$ are visited infinitely often (under stationary fully randomized policy δ) as $t \to \infty$, standard results from stochastic approximation theory [16] can be used to show that $Q_t \to Q^\delta$ almost surely.

All that is required of the non-negative step size parameters $\gamma_t(i,a)$ (for $(i,a) \in \mathcal{Q}$) is that they should satisfy the standard assumptions

$$\sum_{t=1}^{\infty} \gamma_t(i,a) = \infty, \quad \sum_{t=1}^{\infty} (\gamma_t(i,a))^2 < \infty,$$

almost surely and may be allowed to be random and can depend on the past history (at the time the step sized is used).

Consider the following algorithm.

Algorithm 5.5

Let $\underline{0} < \bar{\epsilon}_k < \tilde{\epsilon}$ be a sequence of positive vectors in $\mathbf{R}^n$; where $\tilde{\epsilon} \in \mathbf{R}^n$ is defined as a vector with $\tilde{\epsilon} = \frac{1}{|\mathcal{A}(i)|}$.

1. *Set $k = 0$.*

2. *Select an arbitrary stationary deterministic policy $\mu_0 \in \Upsilon$.*

3. *Choose the stationary fully randomized extremal policy $\delta_k \in \Lambda_{\bar{\epsilon}_k}$, associated with μ_k and run Algorithm 5.4, for "large" random number n_k of steps, till*

Φr_{n_k} *"nearly" converges to* J_{δ_k} *and* Q_{n_k} *"nearly" converges to* Q^{δ_k}*. Let* $\tilde{Q} = Q_{n_k}$ *and* $J_k = \Phi r_{n_k}$*. Let*

$$\zeta_k = \max\{\max_{i \in \mathcal{S}} |J_k(i) - J_{\delta_k}(i)|, \max_{(i,a) \in \mathcal{Q}} |\tilde{Q}(i,a) - Q^{\delta_k}(i,a)|\}.$$

4. *Set* $k = k + 1$ *and update the policy to* μ_k*, where*

$$\mu_k(i) = \arg\min_{a \in \mathcal{A}(i)} \tilde{Q}(i,a)$$

5. *Go to step 3.*

□

Note that the initial values $z_{-1}, \vartheta_0, r_0, i_0$ and Q_0 when calling Algorithm 5.4 at each iteration in step 3 of Algorithm 5.5 may be arbitrary, but may be set to the final values obtained while running Algorithm 5.4 by calling it in step 3 in the previous iteration of Algorithm 5.5.

Theorem 5.3 *Consider the Algorithm 5.5, and let Assumption 5.6 and Assumption 5.8 hold. Then we have the following results.*

1. *Given any scalar* $\varepsilon > 0$*, there exists scalar* $\zeta > 0$ *and vector* $\bar{\epsilon}$*, with* $\underline{0} < \bar{\epsilon} < \tilde{\epsilon}$*, such that if* $\bar{\epsilon}_k < \bar{\epsilon}$ *and* $\zeta_k < \zeta$*,* $\forall k$*; then* μ_k *"converges" to an optimal stationary deterministic policy (if there are multiple stationary deterministic policies that are optimal, it may take any of the stationary deterministic policy which minimizes the RHS of the Bellman Equation 5.8) in a finite number of steps. In particular* J_{μ_k} *converges to* $J^\diamond$ *and* ϑ^{μ_k} *converges to* ϑ^* *in a finite*

number of steps ($\leq |\Upsilon|$). Also

$$|\vartheta^{\delta_k} - \vartheta^{\mu_k}| < \varepsilon, \qquad \textit{and} \quad \max_{i \in \mathcal{S}} (J_{\delta_k}(i) - J_{\mu_k}(i)) < \varepsilon, \quad \forall k.$$

2. *In particular if* $\limsup_{k \to \infty} \bar{\epsilon}_k(i) = 0, \ \forall i \in \mathcal{S}$, *and* $\limsup_{k \to \infty} \zeta_k = 0$, *then* $\| J_{\mu_k} - J^\diamond \| \to 0$, $|\vartheta^{\mu_k} - \vartheta^*| \to 0$, $\| J_{\delta_k} - J^\diamond \| \to 0$ *and* $\vartheta^{\delta_k} \to \vartheta^*$. *Infact* $J_{\mu_k} = J^\diamond$ *and* $\vartheta^{\mu_k} = \vartheta^*$, *for all large* k.

□

Proof of Theorem 5.3

Proof of part (1) follows from Lemma 5.4, Proposition 5.2, Lemma 5.10 and the arguments following Corollary 5.2.

Proof of part (2) follows from part (1) and arguments along the same line as in the proof of part (1).

□

Notice that for communicating MDP which is unichain, the state space $\mathcal{S}$ is the union of the recurrent classes corresponding to the stationary deterministic policies (refer Appendix B). Note that in Algorithm 5.5, we need not know the common recurrent state.

Comments following the proof of Theorem 5.2 hold for Theorem 5.3.

5.9.3 Weakly Communicating Unichain MDP With A Common Recurrent State

Note that any unichain MDP is weakly communicating. We now extend the TD learning scheme to general Unichain MDP with a common recurrent state.

Note that for any stationary fully randomized policy of the Unichain MDP, the unique recurrent class is the union of the recurrent classes of the stationary deterministic policies. Also for any policy, stationary or otherwise, the process almost surely gets absorbed into this unique recurrent class, irrespective of the starting state.

We are interested in finding the states belonging to this unique recurrent class, by simulation. Let $\delta \in \Lambda$ be any stationary randomized policy. The elements of its recurrent class (note that the Markov chain corresponding to δ is unichain) are precisely, those states 'i' for which $\pi^{\delta}(i)$ is positive, where $\pi^{\delta} \in \mathbf{R}^n$ is the unique invariant distribution (or occupation probabilities) of the Markov chain corresponding to δ.

Fix $\delta \in \Lambda$. Let $\gamma_t(i), t \geq 1$ be a sequence of non-negative real valued step size parameters for each $i \in \mathcal{S}$. We are interested in estimating the occupation probability $\pi^{\delta}(i)$, with $U_t(i)$ at time t. We start at time $t = 1$, and let i_t denote the state at time t.

Assumption 5.9 *For each $i \in \mathcal{S}$, the non-negative monotonic step sizes $\gamma_t(i) \downarrow 0$, are pre-determined and satisfy $\sum_{t=1}^{\infty} \gamma_t(i) = \infty$; $\sum_{t=1}^{\infty} \gamma_t^2(i) < \infty$.*

□

We have the following stochastic small step algorithm for estimating the occupation probabilities. The starting state i_1 can be arbitrary.

Algorithm 5.6

$$U_0(i) = 0, \quad \forall i \in \mathcal{S}$$

The update rule is as follows (starting at $t = 1$):

$$U_t(i) = (1 - \gamma_t(i))U_{t-1}(i) + \gamma_t(i)\boldsymbol{\mathcal{I}}_{[i_t=i]}, \quad \forall i \in \mathcal{S}$$

□

Here $\boldsymbol{\mathcal{I}}_{[i_t=i]}$ is the indicator function which takes the value one, if i_t equal to i, and takes the value zero, if i_t not equal to i. Note that if $\gamma_t(i) = \frac{1}{t}$, then $U_t(i)$ is the fraction of time the Markov chain has been in state i, by time t.

Lemma 5.24 *Fix a stationary randomized policy $\delta \in \Lambda$, for the Unichain MDP. Under Assumption 5.9, in Algorithm 5.6, $U_t(i) \to \pi^\delta(i)$ almost surely for each $i \in \mathcal{S}$.*

□

Proof of Lemma 5.24

The proof is straightforward and follows from Theorem 3.2 of Chapter 3.

□

In general $U_0(i)$ can be arbitrary and need not be zero. Also Assumption 5.9 may be replaced by the condition that for each $i \in \mathcal{S}$, the predetermined non-

negative step sizes $\gamma_t(i), t \geq 1$, satisfies $\sum_{t=1}^{\infty} |\gamma_t(i) - \gamma_{t+1}(i)| < \infty$, $\sum_{t=1}^{\infty} \gamma_t(i) = \infty$ and $\sum_{t=1}^{\infty} \gamma_t^2(i) < \infty$.

Hence if we fix a stationary fully randomized policy $\delta \in \Lambda$, (for example the one in which for each state, all the feasible actions are taken with equal probability) and run the above algorithm for sufficiently large t, then we can find the states in the unique recurrent class corresponding to stationary fully randomized policies (they are precisely the states for which $\pi^{\delta}(i) > 0$).

Once we identify this unique recurrent class, we can focus our attention on this subset of states and apply the average cost TD(λ) algorithm developed in subsection 5.9.2 for the MDP restricted to this set of states. Actually for the original MDP, the actions taken at the transient states are irrelevant, since given any policy (stationary or otherwise) the system gets absorbed into the above mentioned unique recurrent class of states, almost surely.

But we might like to solve the Bellman Equation 5.8, for the original MDP; that is we need to find $\vartheta^* \in \mathbf{R}$ and $J^{\diamond} \in \mathbf{R}^n$. This can be done in two steps.

Step 1: First we solve the Bellman equation for the MDP restricted to the unique recurrent class (corresponding to stationary fully randomized policies). Assume without loss of generality that the states in this unique recurrent class are the last $n - m$ states of the MDP. That is, the first m states are the transient states under any stationary fully randomized policy. Then ϑ^*, the optimal average cost along with $(J^{\diamond}(m+1), \ldots, J^{\diamond}(n))'$, the differential cost vector, solves the average cost MDP restricted on the unique recurrent class. This may be approximately obtained as in the previous subsection using average cost TD(λ) schemes.

Step 2: Once we solve the first step, we solve a stochastic shortest path problem (using TD(λ) schemes mentioned in Appendix D) to solve the original average cost Bellman Equation 5.8. The details are given next. Consider an n state Stochastic Shortest Path MDP, in which the feasible actions and transition probabilities for the first m states are exactly similar the the original Average Cost MDP, except that the expected value of immediate cost for taking action a from state i is $(g(i,a) - \vartheta^*)$ (for $i \in \{1,\ldots,m\},\ a \in \mathcal{A}(i)$). For the last $n-m$ states, we assume that it has only one fictitious feasible action from each of these states, under which the system moves to the terminal state 0, with probability one. The corresponding immediate cost for taking this fictitious action from state i being $J^\diamond(i)$, a deterministic quantity (for $i \in \{m+1,\ldots,n\}$). See that for this Stochastic Shortest Path problem (SSP), all stationary deterministic policies are proper. Note that the Bellman Equation for this new SSP [12, 16] is $J^\diamond = \tilde{T}J^\diamond$, where $\tilde{T}$ is the dynamic programming operator for the SSP (see Appendix D). Note that the minimizing action in the Bellman Equation for the SSP is the same as the minimizing action for the Bellman Equation 5.8, for the average cost MDP for states $1,\ldots,m$.

Thus once we have an estimate for ϑ^* and $J^\diamond(m+1),\ldots,J^\diamond(n)$ from step one, we may plug in these estimates for solving the SSP mentioned in step 2, using the TD(λ) schemes in Appendix D.

Note that throughout this section, the moment condition on immediate costs in Assumption 5.6(c), namely

$$E\left[|g_t|^k | i_t = i, u_t = a\right] < \infty,$$

$\forall i \in \mathcal{S}, a \in \mathcal{A}(i)$, need be satisfied only upto a sufficiently large k ($k = 4$) and not for all $k > 0$. (see Chapter 3, Proposition 3.1).

Chapter 6

Conclusion

In Chapter 2 of the dissertation we prove the Lipschitz continuity of the cost go function for the finite horizon and infinite horizon discounted cost POMDP (with domain the unit simplex of probability distributions over the underlying states) and give bounds for the Lipschitz constant. We use these Lipschitz constant bounds to provide error bounds for computational algorithms which rely on the discretization of the unit simplex.

For the computational schemes for POMDPs discussed in Chapter 2, partitioning of the unit simplex Δ_n, and representative points in each member of the partition may be obtained as in Appendix A, by mapping each point in the unit simplex Δ_n to the nearest point in Δ_n^m (the set of representative points), ties being resolved consistently. Larger the value of m, finer the partition.

In Chapter 3 we discuss generalization of a standard stochastic approximation algorithm to handle periodicity of the underlying Markov process. This result is used to extend the proof of convergence of temporal difference (TD) schemes with linear function approximation to estimate cost to go function for discounted cost criterion and differential cost function for average cost criterion. This is an extension of the work in [54, 55].

In Chapter 4 we outline an approximate policy iteration scheme for infinite

horizon discounted cost MDPs, using TD schemes to evaluate the cost to go function for stationary fully randomized policies which are "near" to stationary deterministic policies. This allows for exploration and the corresponding Q-values are estimated via online small step stochastic approximation and this in turn is used for policy improvement.

In Chapter 5 we outline an approximate policy iteration scheme for average cost unichain MDPs with a common recurrent state. We use TD schemes to evaluate the differential cost for stationary policies. Corresponding Q-values are also estimated (incorporating exploration using stationary randomized policies which are "near" to stationary deterministic policies) via online small step stochastic approximation and this in turn is used for policy improvement.

Appendix A deals with a discretization scheme for unit simplex.

Appendix B deals with reachability structure of finite state finite action MDPs.

Appendix C deals with error bounds for MDPs and some contraction mapping theorems.

Appendix D deals with TD schemes for stochastic shortest path MDPs.

Some notes on the discounted cost MDP and average cost MDP follow.

We may use either the Least Squares Policy Evaluation (LSPE(λ)) or the Least squares Temporal Difference (LSTD(λ)) [12] for policy evaluation instead of the TD(λ) schemes in the discounted cost and average cost cases of Chapter 4 and Chapter 5 respectively. These schemes converge much faster than the TD(λ) schemes, but is computationally expensive at each step. For example, if we use a basis of K vectors for linear function approximation, at each time step we need the

inversion of a $K \times K$ matrix and a fixed number of matrix vector multiplications (matrix multiplying a vector) in both the LSPE(λ) and LSTD(λ). Multiplication of a vector of dimension K with a $K \times K$ matrix involves computation of order $\mathcal{O}(K^2)$. Inversion of a $K \times K$ matrix is an order $\mathcal{O}(K^3)$ operation. Howerver because of the particular way in which the matrix evolves, we may use the matrix inversion lemma and can manage to compute the matrix inversion with $\mathcal{O}(K^2)$ operation. Hence the LSPE(λ) or LSTD(λ) requires $\mathcal{O}(K^2)$ operation at each step, whereas TD(λ) needs only $\mathcal{O}(K)$ operation per time step.

Note that for the discounted cost MDP, if we define, for each stationary deterministic policy $\mu \in \Upsilon$, the "greedy region" for μ as

$$\mathcal{R}_\mu = \{J \in \mathbf{R}^n | T_\mu J = TJ\},$$

where the operators T and T_μ are defined as in Chapter 4, then $\mathcal{R}_\mu$ is a polyhedron. It might be an empty set. In Chapter 4, if we use linear function approximation (instead of look up table representation) for the TD(λ) schemes for discounted cost problems, along with approximate policy iteration, and the space spanned by the basis functions does not intersect the greedy region corresponding to any of the optimal stationary deterministic policies, then the methodology in Chapter 4 cannot converge to an optimal policy.

The same observations hold even if we use TD(λ) schemes for discounted cost problems, along with approximate policy iteration, using the equivalent Stochastic Shortest Path (SSP) formulation [16].

For the average cost MDP, we may define for each stationary deterministic

policy $\mu \in \Upsilon$, the "greedy region" for μ as

$$\bar{\mathcal{R}}_\mu = \{J \in \mathbf{R}^n | \bar{T}_\mu J = \bar{T} J\},$$

where $\bar{T}_\mu$ and $\bar{T}$ are defined as in Chapter 5. Here also $\bar{\mathcal{R}}_\mu$ is a polyhedron and may be an empty set. In Chapter 5, if we use linear function approximation (instead of look up table representation) for the TD(λ) schemes for average cost problem, along with approximate policy iteration, and the space spanned by the basis functions does not intersect the "greedy region" corresponding to any of the optimal stationary deterministic policies, then the methodology in Chapter 5 cannot converge to an optimal policy.

The above observations also hold when we use linear function approximation with LSPE(λ) and LSTD(λ) policy evaluation.

For a comparitive study of discounted versus average cost temporal difference schemes for a Markov Cost Process see [56].

For optimistic policy iteration schemes see [16, 52].

6.1 Future Work : Extension Of Reinforcement Learning To POMDPs

We consider direct adaptive control of POMDPs in this section for the infinite horizon discounted cost case (with discount factor $\beta \in [0, 1)$) using temporal difference schemes to obtain near optimal policies. See Chapter 2 for details about the definition of a POMDP. We need not know about the cardinality of the underlying state space, which is assumed finite. We need to know the finite set of feasible actions $\mathcal{A}$, common to all underlying states. We also need to know the finite set of

observations, namely $\mathcal{O}$. We need to observe the immediate cost incurred as well as the associated observation, in response to taking an action, at each time step. We don't assume any direct knowledge about the underlying probabilities, but assume that the immediate costs have finite moments for the underlying MDP. We assume that the underlying MDP is communicating.

Fix an integer $N > 0$. Consider an *associated* MDP with state at time t, given by the tuplet $(s_{t-N+1}, o_{t-N+1}, u_{t-N+1}, s_{t-N+2}, o_{t-N+2}, u_{t-N+2}, \cdots, s_{t-1}, o_{t-1}, u_{t-1}, s_t, o_t)$, where s_t is the underlying state at time t, of the original MDP, u_t is the action taken at time t, o_t is the observation obtained at time t in response to the action u_{t-1} taken at time $t-1$ and subsequent transition of the underlying state from s_{t-1} to s_t.

Note that the observation o_t at time t, might also include a finite discretized version of g_{t-1}, along with the traditional observation from the finite set $\mathcal{O}$. This is because the immediate cost incurred at time $t-1$ (in response to taking action u_{t-1}), namely g_{t-1}, might contain information about the underlying state s_t of the MDP.

The feasible action set for the associated MDP is same as that of the original POMDP. The transition probability and the immediate cost for the associated MDP is obtained in the most natural way (we omit the details) from the original POMDP. At time t, the N-stage observable history is $(o_{t-N+1}, u_{t-N+1}, \cdots, o_{t-1}, u_{t-1}, o_t)$, and is considered to be the *pseudo state* at time t, and denoted by $\tilde{s}_t$. When $N = 1$, the pseudo state at time t is just (o_t).

We work with stationary fully randomized policies with the pseudo state as the "current state". Such a policy is also a stationary fully randomized policy on

the associated MDP mentioned earlier.

Once we fix a stationary fully randomized policy δ, we may use any linear function approximation (on the pseudo states) along with TD(λ) (as in Chapter 4), to obtain an estimate of the approximate cost to go $\tilde{J}^{\delta}(\tilde{s})$ from the pseudo state $\tilde{s}$. The restriction on the step sizes γ_t used in the TD(λ) scheme is as in Chapter 4. We call any collection of pseudo states an *aggregated pseudo state* $\hat{s}$. We may estimate the Q-value corresponding to any "aggregated pseudo state - action" pair (with one step look ahead function $\tilde{J}^{\delta}$) by the small step stochastic approximation

$$Q_{t+1}(\hat{s},u) = (1-\gamma_{\tau_t(\hat{s},u)}(\hat{s},u))Q_t(\hat{s},u) + \gamma_{\tau_t(\hat{s},u)}(\hat{s},u)\left[g_t + \beta\tilde{J}_{t+1}(\tilde{s}_{t+1})\right]$$

if and only if $\tilde{s}_t \in \hat{s}$ and $u_t = u$; otherwise $Q_{t+1}(\hat{s},u) = Q_t(\hat{s},u)$. Here $\beta \in [0,1)$ is the discount factor, $\hat{s}$ is the aggregated pseudo state, and u is the action under consideration. $\tilde{J}_t(\tilde{s}_t)$ is the estimate of $\tilde{J}^{\delta}(\tilde{s}_t)$ at time t obtained via TD(λ) scheme. In the above update equation for Q_{t+1}, we could have used $\tilde{J}_t(\tilde{s}_{t+1})$ instead of $\tilde{J}_{t+1}(\tilde{s}_{t+1})$. Here $\tau_t(\hat{s},u)$ is the number of times action u is taken by time t, from any of the pseudo states belonging to the aggregated pseudo state $\hat{s}$. Here for $\hat{s}$ and u, $\gamma_k(\hat{s},u)$ is a deterministic non-negative step size sequence which satisfies

$$\sum_{k=1}^{\infty}\gamma_k(\hat{s},u) = \infty; \quad \sum_{k=1}^{\infty}\gamma_k^2(\hat{s},u) < \infty; \quad \sum_{k=1}^{\infty}|\gamma_{k+1}(\hat{s},u) - \gamma_k(\hat{s},u)| < \infty.$$

In particular a non-increasing non-negative sequence satisfies the third condition above. We may prove that $Q_t(\hat{s},u)$ converges to a quantity $Q^{\delta}(\hat{s},u)$ (which actually depends also on $\tilde{J}^{\delta}(\cdot)$ and hence on the choice of the basis functions for linear function approximation) which may be characterized analytically (we omit the details).

In particular, we may partition the space of pseudo states, and use the indicator functions for each member of the partition (the aggregated pseudo state) as the basis functions for the linear approximation, and correspondingly estimate the cost to go and Q-values for aggregated state - action pairs. Here each member of the partition is considered to be an aggregated pseudo state. But does these Q-values and cost to go approximation converge to anything useful?

Suppose the original POMDP is such that the aposteriori probability on the underlying states, given the past observable history becomes less and less dependent on the initial apriori probability on the underlying states, uniformly for all observable past histories; then for sufficiently large fixed integer N, each pseudo state corresponds roughly to a point (actually a small neighbourhood of a point) on the unit simplex of belief states (the space of probability distributions on the underlying states of the original POMDP).

If we partition the space of pseudo states, such that each member of the partition (aggregated pseudo state) has the property that the pseudo state in each aggregated pseudo state corresponds roughly to the same point (or neighbourhood) of the unit simplex of the belief states, then the TD(λ) scheme for learning as in Chapter 4 leads to a near optimal solution for the original POMDP.

In this scheme, only those belief states (actually neighbourhoods) which correspond to the N-stage observable history are involved or explored, which is all what we need.

Again we could have used LSPE(λ) or LSTD(λ) instead of TD(λ) to estimate the cost to go .

Appendix A

Discretization Of The Unit Simplex

We are interested in the simple problem of approximation (by discretization) of probability mass functions on a finite sample space. Please note that the notations in this appendix are self contained. For any positive integer n, let

$$\Delta_n \equiv \left\{ p = (p_1, p_2, \ldots, p_n) \mid p_i \geq 0,\ i = 1, \ldots, n;\ \sum_{i=1}^{n} p_i = 1 \right\}$$

be the $n-1$ dimensional unit simplex in $\mathbf{R}^n$ (the n dimensional Eucledian space).

For any positive integers m and n, let

$$\Delta_n^m \equiv$$
$$\left\{ q = (q_1, q_2, \ldots, q_n) \mid q_i = \frac{l_i}{m},\ l_i \text{ non-negative integer}, i = 1, \ldots, n; \sum_{i=1}^{n} l_i = m \right\}.$$

For any positive integers m, n and a non-negative integer l with $0 \leq l \leq m$, define

$$\Delta_n^{m,l} \equiv$$
$$\left\{ v = (v_1, v_2, \ldots, v_n) \mid v_i = \frac{l_i}{m},\ l_i \text{ non-negative integer}, i = 1, \ldots, n; \sum_{i=1}^{n} l_i = m - l \right\}.$$

Note that $\Delta_n^m \subset \Delta_n$ and $\Delta_n^{m,0} = \Delta_n^m$. The cardinality of the set $\Delta_n^{m,l}$ is $|\Delta_n^{m,l}| = \frac{(m-l+n-1)!}{(m-l)!\,(n-1)!}$.

For any real $\alpha \in [1, +\infty)$ we define the ℓ_α norm as follows,

$$\|v\|_\alpha = \left(\sum_{i=1}^{n} |v_i|^\alpha \right)^{\frac{1}{\alpha}}, \qquad v \in \mathbf{R}^n$$

and the ℓ_∞ norm as

$$\|v\|_\infty = \max_{i\in\{1,2,\ldots,n\}} |v_i|, \qquad v \in \mathbf{R}^n.$$

We have the following lemma.

Lemma A.1 *For any $v \in \mathbf{R}^n$, $\|v\|_\alpha \downarrow \|v\|_\infty$ as $\alpha \to +\infty$.*

□

Proof of Lemma A.1

If $v_i = 0,\ \forall i \in \{1, 2, \ldots, n\}$, i.e. $v = \underline{0}$ ($\underline{0}$ is the zero vector), then the claim is trivially true. Suppose $v \neq \underline{0}$. Without loss of generality assume that $|\, v_i \,| \geq |\, v_{i+1} \,|,\ i = 1, 2, \ldots, n-1$. Otherwise we could do a permutation of the indices without changing the norm. Note that $|v_1| > 0$. Let $\tilde{v}_i = |v_i| \,/\, |v_1|$. Note that $0 \leq \tilde{v}_i \leq 1,\ \forall i$ with $\tilde{v}_1 = 1$. Then $\|v\|_\infty = |v_1|$ and $\|v\|_\alpha = |v_1|\ (\sum_{i=1}^n \tilde{v}_i^\alpha)^{\frac{1}{\alpha}}$ for $\alpha \in [1, +\infty)$. If $1 \leq \alpha < \beta < +\infty$, then $1 \leq (\sum_{i=1}^n \tilde{v}_i^\beta) \leq (\sum_{i=1}^n \tilde{v}_i^\alpha) \leq n$. This implies that $1 \leq (\sum_{i=1}^n \tilde{v}_i^\beta)^{\frac{1}{\beta}} \leq (\sum_{i=1}^n \tilde{v}_i^\beta)^{\frac{1}{\alpha}} \leq (\sum_{i=1}^n \tilde{v}_i^\alpha)^{\frac{1}{\alpha}} \leq n^{\frac{1}{\alpha}}$. This proves that $\|v\|_\beta \leq \|v\|_\alpha$ and $\|v\|_\alpha \downarrow \|v\|_\infty$ as $\alpha \to +\infty$.

□

Fix positive integers n, m. Given a $p \in \Delta_n$, we are interested in finding an element $q \in \Delta_n^m$ depending on p, such that the metric $\|p - q\|_1$ is minimized. We seek to find a function $f : \Delta_n \to \Delta_n^m$ such that

$$\|f(p) - p\|_1 = \inf_{\hat{q}\in\Delta_n^m} \|\hat{q} - p\|_1$$

f need not be a true function in the sense that for a given argument $p \in \Delta_n$, it can

pick any element $q \in \Delta_n^m$ (if there are ties), such that $\|p - q\|_1$ is the minimum. Since the case $n = 1$ is trivial, we assume $n > 1$.

For a real number β, $\lfloor \beta \rfloor$ denotes the floor of β. Consider the following algorithm to find an f.

Algorithm A.1

1. *Given any* $p \in \Delta_n$, *let* $w \in \mathbf{R}^n$ *be chosen such that* $w_i = \lfloor (p_i\, m) \rfloor \frac{1}{m}$; $i = 1, \ldots, n$. *Note that* $0 \leq p_i - w_i < \frac{1}{m}$.

   Define the function $g : \Delta_n \to \bigcup_{l=0}^{m} \Delta_n^{m,l}$ *by* $g(p) = w$.

2. *Let* $k = m\,(1 - (\sum_{i=1}^{n} w_i))$. *Note that* k *is an integer such that* $0 \leq k \leq \min\{m, n-1\}$.

   Define the function $h : \Delta_n \to \{0, 1, \ldots, m\}$ *by* $h(p) = k$.

3. *Order the* n *indices into* $i_1, i_2, \ldots, i_n$, *such that* $(p_{i_j} - w_{i_j}) \geq (p_{i_{j+1}} - w_{i_{j+1}})$, $j = 1, \ldots, n-1$. *Ties may be resolved arbitrarily.*

4. *If* $k = 0$, *then set* $\hat{k} = 0$, *else set* $\hat{k} = \max\{j \,|\, (p_{i_j} - w_{i_j}) > 0\}$. *If* $k = 0$ *then set* $\tilde{k} = 0$, *else set* $\tilde{k} = \max\{j \mid k \leq j \leq n \text{ and } (p_{i_j} - w_{i_j}) = (p_{i_k} - w_{i_k})\}$.

5. *Let* $q \in \mathbf{R}^n$ *be defined by the folowing steps.*

   - *For* $j = k+1, \ldots, n$, *set* $q_{i_j} = w_{i_j}$.
   - *If* $k > 0$,*then for* $j = 1, \ldots, k$, *set* $q_{i_j} = w_{i_j} + \frac{1}{m}$.

   Note that $q \in \Delta_n^m$.

6. Set $f(p) = q$.

□

Remarks on Algorithm A.1

Note that f depends on n and m. Observe that $g(p) \in \Delta_n^{m,k}$ and $h(p) \in \{0, 1, \cdots, \min\{m, n-1\}\}$ are well defined. Note that $f(p) = p$ if and only if $k = 0$ in step 2 of Algorithm A.1. Note that $\|f(p) - p\|_\infty < \frac{1}{m}$. Also note that $\hat{k} > k$ whenever $k > 0$, since $\sum_{i=1}^{n}(q_i - p_i) = 0$. We have $k \leq \tilde{k} \leq \hat{k}$. When $k > 0$, we have $\frac{1}{m} > (q_{i_j} - p_{i_j}) > 0$ for $j = 1, \ldots, k$; $-\frac{1}{m} < (q_{i_j} - p_{i_j}) < 0$ for $j = k+1, \ldots, \hat{k}$ and $p_{i_j} = q_{i_j}$ for $\hat{k} < j \leq n$. If $k > 1$, then $0 < (q_{i_j} - p_{i_j}) \leq (q_{i_{j+1}} - p_{i_{j+1}})$ for $j = 1, \ldots, k-1$. Also if $\hat{k} > k+1$, then $(p_{i_j} - q_{i_j}) \geq (p_{i_{j+1}} - q_{i_{j+1}}) > 0$ for $j = k+1, \ldots, \hat{k}-1$. If $p = (0, 0, \ldots, 0, 1, 0, \ldots, 0, 0)$, with an entry one in the i^{th} position, and zero elsewhere, then $k = 0$ in step 2 of Algorithm A.1. Also note that for any fixed $\alpha \in [1, +\infty]$, $\|f(p) - p\|_\alpha$ is the same irrespective of the ordering taken in step 3 of Algorithm A.1, when ties need to be resolved arbitrarily.

□

Fix positive integers n, m. For any fixed $p \in \Delta_n$, define

$$C(p) \equiv \left\{ \tilde{q} \in \Delta_n^m \mid \|p - \tilde{q}\|_\infty < \frac{1}{m} \right\}.$$

Note that this is a non-empty set ($f(p) \in C(p)$). Also note that any element in $C(p)$ is obtained as follows. If $f(p) = p$ then $C(p) = \{p\}$, a singleton set. Otherwise (i.e. if $k > 0$ in step 2 of Algorithm A.1) we have

$$C(p) = \left\{ \tilde{q} \in \Delta_n \,|\, \tilde{q}_i = w_i + \frac{1}{m}, i \in E, \tilde{q}_i = w_i, i \notin E; \text{ for some } E \in D(p) \right\}.$$

Here $w \in \mathbf{R}^n$ is obtained from step 1 of Algorithm A.1, and $D(p) \equiv \{E \subset \{i_1, i_2, \ldots, i_{\hat{k}}\} \mid \mid E \mid = k\}$. Note that the indices $i_1, i_2, \ldots, i_{\hat{k}}$ are obtained from steps 3 and 4 in Algorithm A.1 and depend only on p, n and m. Here $|E|$ represents the cardinality of the set E.

We state and prove the following result.

Lemma A.2 *The function f defined in Algorithm A.1 satisfies*

$$\|f(p) - p\|_\alpha = \inf_{\tilde{q} \in \Delta_n^m} \|\tilde{q} - p\|_\alpha$$

for any $p \in \Delta_n$ and any $\alpha \in [1, +\infty]$.

□

Proof of Lemma A.2

We will show that $\tilde{q} \in \Delta_n^m \setminus C(p)$ implies that there exists $\hat{q} \in \Delta_n^m$ such that $\|p - \hat{q}\|_\alpha < \|p - \tilde{q}\|_\alpha$. This is easy for the case $\alpha = +\infty$, since $\hat{q} = f(p)$ does the job. Suppose $\alpha \in [1, \infty)$. Then since $\tilde{q} \in \Delta_n^m \setminus C(p)$, there exists an index i such that $\|p - \tilde{q}\|_\infty = |p_i - \tilde{q}_i| \geq \frac{1}{m}$. Then either $p_i \geq \tilde{q}_i + \frac{1}{m}$ or $p_i \leq \tilde{q}_i - \frac{1}{m}$.

First we consider the case $p_i \geq \tilde{q}_i + \frac{1}{m}$. Then there exists index j such that $p_j < \tilde{q}_j$ since $\sum_{l=1}^n (p_l - \tilde{q}_l) = 0$. Define $\hat{q} \in \Delta_n^m$ as follows : $\hat{q}_l = \tilde{q}_l, l \neq i, l \neq j$, $\hat{q}_i = \tilde{q}_i + \frac{1}{m}$ and $\hat{q}_j = \tilde{q}_j - \frac{1}{m}$. Note that $\hat{q} \in \Delta_n^m$. We will show that $|p_i - \tilde{q}_i|^\alpha + |p_j - \tilde{q}_j|^\alpha > |p_i - \hat{q}_i|^\alpha + |p_j - \hat{q}_j|^\alpha$ and $\|p - \tilde{q}\|_\alpha > \|p - \hat{q}\|_\alpha$ follows. Let $x \equiv p_i - \tilde{q}_i$. Then $|p_i - \tilde{q}_i| = x \geq \frac{1}{m}$ and $x > |p_i - \hat{q}_i| = p_i - \hat{q}_i = x - \frac{1}{m} \geq 0$. Supposing that $\tilde{q}_j \geq p_j + \frac{1}{m}$, it is clear that $|p_j - \tilde{q}_j| > |p_j - \hat{q}_j| = (\hat{q}_j - p_j) = (\tilde{q}_j - p_j - \frac{1}{m}) \geq 0$ and hence $|p_i - \tilde{q}_i|^\alpha + |p_j - \tilde{q}_j|^\alpha > |p_i - \hat{q}_i|^\alpha + |p_j - \hat{q}_j|^\alpha$. On the other hand, if

$0 < (\tilde{q}_j - p_j) < \frac{1}{m}$, then $y \equiv p_j - \hat{q}_j$ is such that $0 < y < \frac{1}{m}$. We show that

$$x^\alpha + (\frac{1}{m} - y)^\alpha > (x - \frac{1}{m})^\alpha + y^\alpha$$

for $\alpha \in [1, +\infty)$. Note that $(\frac{1}{m})^\alpha > ((\frac{1}{m} - y)^\alpha - y^\alpha) > -(\frac{1}{m})^\alpha$. Hence all that is required to prove the above inequality is to show that $x^\alpha - (x - \frac{1}{m})^\alpha \geq (\frac{1}{m})^\alpha$. Now $x^\alpha(1-(1-\frac{1}{m}\frac{1}{x})^\alpha) \geq x^\alpha(\frac{1}{m}\frac{1}{x})^\alpha$, since for $0 \leq \beta \leq 1$ and $\alpha \in [1, +\infty)$, we have $1-\beta^\alpha \geq (1-\beta)^\alpha$. Thus we have shown that $|p_i - \tilde{q}_i|^\alpha + |p_j - \tilde{q}_j|^\alpha > |p_i - \hat{q}_i|^\alpha + |p_j - \hat{q}_j|^\alpha$.

For the case where $p_i \leq \tilde{q}_i - \frac{1}{m}$, note that there exists an index j such that $p_j > \tilde{q}_j$ since $\sum_{l=1}^n (p_l - \tilde{q}_l) = 0$. We define $\hat{q} \in \Delta_n^m$ as follows : $\hat{q}_l = \tilde{q}_l, l \neq i, l \neq j$, $\hat{q}_i = \tilde{q}_i - \frac{1}{m}$ and $\hat{q}_j = \tilde{q}_j + \frac{1}{m}$. Now an argument similar to that in the above paragraph shows that $|p_i - \tilde{q}_i|^\alpha + |p_j - \tilde{q}_j|^\alpha > |p_i - \hat{q}_i|^\alpha + |p_j - \hat{q}_j|^\alpha$ and hence $\|p - \tilde{q}\|_\alpha > \|p - \hat{q}\|_\alpha$.

Now to prove our main result, namely $\| f(p) - p\|_\alpha = \inf_{\tilde{q} \in \Delta_n^m} \| \tilde{q} - p\|_\alpha$, we use an argument of contradiction. We consider the cases $\alpha \in [1, \infty)$ and $\alpha = +\infty$ separately.

First assume that $\alpha \in [1, +\infty)$. Suppose there exists $\tilde{q} \in \Delta_n^m$ such that $\inf_{\tilde{q} \in \Delta_n^m} \| \tilde{q} - p\|_\alpha = \|\tilde{q} - p\|_\alpha < \|f(p) - p\|_\alpha$. We need to consider only that case in which $\|f(p) - p\|_\alpha > 0$. By way of the argument in the earlier paragraphs, $\tilde{q} \in C(p)$. Since $\tilde{q} \neq f(p)$, we have indices $i \neq j$ and real values x and y with $\frac{1}{m} > x > y > 0$ such that $\tilde{q}_i - p_i = \frac{1}{m} - y > 0$ and $p_j - \tilde{q}_j = x > 0$. Now consider the vector $\hat{q} \in \Delta_n^m$ defined as follows $\hat{q}_l = \tilde{q}_l, l \neq i, l \neq j$, $\hat{q}_i = \tilde{q}_i - \frac{1}{m}$ and $\hat{q}_j = \tilde{q}_j + \frac{1}{m}$. It is easy to see that $\hat{q} \in C(p)$. We will show that for $\alpha \in [1, +\infty)$, $\|\hat{q} - p\|_\alpha < \|\tilde{q} - p\|_\alpha$. In fact, it is sufficient to show that $|p_i - \hat{q}_i|^\alpha + |p_j - \hat{q}_j|^\alpha < |p_i - \tilde{q}_i|^\alpha + |p_j - \tilde{q}_j|^\alpha$. That is, we need to show that $y^\alpha + (\frac{1}{m} - x)^\alpha < (\frac{1}{m} - y)^\alpha + x^\alpha$, which is true by virtue of the

fact that $x^\alpha > y^\alpha$ and $(\frac{1}{m} - y)^\alpha > (\frac{1}{m} - x)^\alpha$, since $\frac{1}{m} > x > y > 0$. This along with the fact that for any fixed $\alpha \in [1, +\infty]$, $\|f(p) - p\|_\alpha$ is the same irrespective of the ordering taken in step 3 of Algorithm A.1, when ties need to be resolved arbitrarily, implies that $\|f(p) - p\|_\alpha = \inf_{\tilde{q} \in \Delta_n^m} \|\tilde{q} - p\|_\alpha$.

Next we consider the case when $\alpha = +\infty$. Fix a $p \in \Delta_n$ and any $\tilde{q} \in \Delta_n^m$. For any $\alpha \in [1, +\infty)$ we have $\|f(p) - p\|_\alpha \leq \|\tilde{q} - p\|_\alpha$. Since for any $v \in \mathbf{R}^n$, $\|v\|_\alpha \downarrow \|v\|_\infty$ as $\alpha \uparrow +\infty$, we have $\|f(p) - p\|_\infty \leq \|\tilde{q} - p\|_\infty$. Since this is true for all $\tilde{q} \in \Delta_n^m$ we have $\|f(p) - p\|_\infty = \inf_{\tilde{q} \in \Delta_n^m} \|\tilde{q} - p\|_\infty$.

□

Fix any $p \in \Delta_n$ and let $\tilde{q} \in \Delta_n^m$. Consider the case when $\alpha \in [1, +\infty)$. Then the proof of the above theorem shows that $\|p - \tilde{q}\|_\alpha = \inf_{\check{q} \in \Delta_n^m} \|\check{q} - p\|_\alpha$ if and only if $\tilde{q} = f(p)$ where f is obtained by Algorithm A.1. Now consider the case when $\alpha = +\infty$. $\|p - \tilde{q}\|_\infty = \inf_{\check{q} \in \Delta_n^m} \|\check{q} - p\|_\infty$, does not necessarily imply that $\tilde{q}$ is of the form $f(p)$, where f is obtained by Algorithm A.1.

Fix positive integers n, m. We give the following error bounds when $\alpha = 1$.

Lemma A.3 *When $\alpha = 1$ we have the following bounds on the approximation error*

$$\begin{aligned}
\sup_{p \in \Delta_n} \|f(p) - p\|_1 &= 2\frac{(n-m)}{n} && \text{if} \quad m < \left\lfloor \frac{n}{2} \right\rfloor \\
&= \frac{1}{2}\frac{(n^2-1)}{m}\frac{1}{n} && \text{if} \quad n \text{ is odd and } m \geq \left\lfloor \frac{n}{2} \right\rfloor \\
&= \frac{1}{2}\frac{n}{m} && \text{if} \quad n \text{ is even and } m \geq \left\lfloor \frac{n}{2} \right\rfloor
\end{aligned}$$

□

Proof of Lemma A.3

Fix positive integers n, m. First of all we define $\mathcal{H}(k) \equiv \{\hat{p} \in \Delta_n \mid h(\hat{p}) = k\}$ for $k = 0, 1, \cdots, \min\{m, n-1\}$ (refer to Algorithm A.1 for the definition of $h(\cdot)$ and $g(\cdot)$). Observe that $\mathcal{H}(k) = \{\hat{p} \in \Delta_n \mid g(\hat{p}) \in \Delta_n^{m,k}\}$.

Fix a $p \in \Delta_n$. Let the corresponding w_i, $i = 1, \ldots, n$ be obtained as in Algorithm A.1 (i.e. $g(p) = w$) and the corresponding ordering of indices be $i_1, i_2, \ldots, i_n$. Let k, $\hat{k}$ and $f(p)$ be obtained as in Algorithm A.1. This implies that $p \in \mathcal{H}(k)$.

When $k = 0$ we have $\|p - f(p)\|_1 = 0$. We focus on the case when $k > 0$ (note that $k \leq n-1$). Let $a_{i_j} \equiv p_{i_j} - w_{i_j} \;\; j = 1, \ldots, n$. See that $\frac{1}{m} > a_{i_1} \geq a_{i_2} \geq \cdots \geq a_{i_n} \geq 0$. Also see that $a_{i_{\hat{k}}} > 0$. When $\hat{k} < n$, we have $a_{i_{\hat{k}+1}} = 0$. Note that $f(p)$ is defined as follows, namely $[f(p)]_{i_j} = w_{i_j} + \frac{1}{m} \;\; j = 1, \ldots, k$ and $[f(p)]_{i_j} = w_{i_j} \;\; j = k+1, \ldots, n$. Let $\varepsilon \in \mathbf{R}$ be such that $k\,\varepsilon = \sum_{j=1}^{k}(\frac{1}{m} - a_{i_j}) = \sum_{j=k+1}^{n} a_{i_j} = \frac{1}{2}\,\|f(p) - p\|_1 > 0$. Note that $(\frac{1}{m} - \varepsilon) \geq a_{i_{k+1}} > 0$ and hence $(n-k)(\frac{1}{m} - \varepsilon) \geq k\,\varepsilon$. This implies $\varepsilon \leq \bar{\varepsilon}$ where $\bar{\varepsilon} \equiv \frac{(n-k)}{n}\,\frac{1}{m}$. Given any such p, we can define a $\tilde{p} \in \Delta_n$ as follows. $\tilde{p}_{i_j} = w_{i_j} + \frac{1}{m} - \varepsilon, \;\; j = 1, \cdots, k$ and $\tilde{p}_{i_j} = w_{i_j} + \frac{k\,\varepsilon}{(n-k)}, \;\; j = k+1, \cdots, n$. Note that $g(\tilde{p}) = w$ (refer to Algorithm A.1) and $(\tilde{p}_{i_j} - w_{i_j}) \geq (\tilde{p}_{i_{j+1}} - w_{i_{j+1}}), \;\; j = 1, \cdots, n-1$. Hence $f(p)$ also serves as $f(\tilde{p})$ and $\|p - f(p)\|_1 = \|\tilde{p} - f(\tilde{p})\|_1$.

Let $\bar{p} \in \Delta_n$ be defined as follows : $\bar{p}_{i_j} = w_{i_j} + \frac{1}{m} - \bar{\varepsilon}, \;\; j = 1, \cdots, k$ and $\bar{p}_{i_j} = w_{i_j} + \frac{k}{n}\frac{1}{m}, \;\; j = k+1, \cdots, n$. Note that $g(\bar{p}) = w$ (refer to Algorithm A.1) and $(\bar{p}_{i_j} - w_{i_j}) \geq (\bar{p}_{i_{j+1}} - w_{i_{j+1}}), \;\; j = 1, \cdots, n-1$. Also $\|f(\bar{p}) - \bar{p}\|_1 = 2k\bar{\varepsilon}$. Hence $f(p)$ also serves as $f(\bar{p})$ and a bit of thought shows that

$$2k\frac{(n-k)}{n}\,\frac{1}{m} \;\; = \|\bar{p} - f(\bar{p})\|_1 = \sup_{\hat{p}:\hat{p} \in \mathcal{H}(k)} \|\hat{p} - f(\hat{p})\|_1$$

Maximizing over k gives the result.

□

Fix positive integers n, m. For any $\tilde{q} \in \Delta_n^m$, define

$$\mathcal{M}(\tilde{q}) \equiv \left\{ \breve{q} \in \Delta_n^m \mid \|\breve{q} - \tilde{q}\|_\infty \leq \frac{1}{m} \right\}.$$

Define $I(\tilde{q}) \equiv \{i \,|\, 0 < \tilde{q}_i < 1\}$ and $J(\tilde{q}) \equiv \{i \,|\, \tilde{q}_i = 0\}$. Let $r(\tilde{q}) \equiv |I(\tilde{q})|$, $s(\tilde{q}) \equiv |\{i \,|\, \tilde{q}_i = 1\}|$ and $t(\tilde{q}) \equiv |J(\tilde{q})|$. Note that $s(\tilde{q})$ is either zero or one and that $r(\tilde{q}) + s(\tilde{q}) + t(\tilde{q}) = n$. If $s(\tilde{q}) = 1$ then $r(\tilde{q}) = 0$ and $t(\tilde{q}) = n - 1$. Note that for any $\hat{q}, \tilde{q} \in \Delta_n^m$, $\|\hat{q} - \tilde{q}\|_\infty = l \frac{1}{m}$ for some integer l, such that $0 \leq l \leq m$, with $\|\hat{q} - \tilde{q}\|_\infty = 0$ iff $\hat{q} = \tilde{q}$. For the set $\mathcal{M}(\tilde{q})$, $|\mathcal{M}(\tilde{q})|$ denotes the cardinality of the set. We have the following lemma.

Lemma A.4 *Let $\tilde{q} \in \Delta_n^m$. Then $|\mathcal{M}(\tilde{q})| = n$ if $s(\tilde{q}) = 1$, else if $s(\tilde{q}) = 0$ then*

$$|\mathcal{M}(\tilde{q})| = \sum_{l=0}^{\lfloor \frac{r(\tilde{q})}{2} \rfloor} \binom{r(\tilde{q})}{l} \binom{n-l}{r(\tilde{q}) - 2l}$$

□

Proof of Lemma A.4

The case when $s(\tilde{q}) = 1$ is trivial. We focus on the case $s(\tilde{q}) = 0$. When $s(\tilde{q}) = 0$, we have $r(\tilde{q}) + t(\tilde{q}) = n$ and $r(\tilde{q}) > 0, t(\tilde{q}) \geq 0$. Note that $I(\tilde{q}) \cup J(\tilde{q}) = \{1, 2, \ldots, n\}$ when $s(\tilde{q}) = 0$. Also $I(\tilde{q}) \cap J(\tilde{q}) = \emptyset$. Let $v \in \mathbf{R}^n$ be defined as $v_i = \tilde{q}_i - \frac{1}{m},\ i \in I(\tilde{q})$ and $v_i = \tilde{q}_i,\ i \in J(\tilde{q})$. Any $\breve{q} \in \mathcal{M}(\tilde{q})$ has the following form, namely $\breve{q}_i = v_i + 2\frac{1}{m},\ i \in A$, $\breve{q}_i = v_i + \frac{1}{m}, i \in B$ and $\breve{q}_i = v_i,\ i \notin A \cup B$. Here $A \subset I(q)$ is such that $0 \leq |A| \leq \lfloor \frac{r(\tilde{q})}{2} \rfloor$, $B \subset \{1, \ldots, n\} \setminus A$ is such that $|B| = r(\tilde{q}) - 2\,|A|$.

Thus the various possible number of ways of choosing $\check{q} \in \mathcal{M}(\tilde{q})$ is

$$|\mathcal{M}(\tilde{q})| = \sum_{l=0}^{\lfloor \frac{r(\tilde{q})}{2} \rfloor} \begin{pmatrix} r(\tilde{q}) \\ l \end{pmatrix} \begin{pmatrix} n-l \\ r(\tilde{q})-2l \end{pmatrix}$$

□

Note that for positive integer k and non-negative integer $l \leq k$,

$$\begin{pmatrix} k \\ l \end{pmatrix} = \frac{k!}{(k-l)!\ l!}.$$

Appendix B

Notes On The Reachability Structure Of Finite State-Finite Action MDP

Consider a finite state-finite action MDP, with state space $\mathcal{S} \equiv \{1, 2, \cdots, n\}$ for some finite integer n. Let the finite non-empty control constraint sets $\mathcal{A}(i) = \{1, 2, \cdots, |\mathcal{A}(i)|\}$ denote the possible control actions from state $i \in \mathcal{S}$. Define $\mathcal{A} = \bigcup_{i=1}^{n} \mathcal{A}(i)$. Let $p_{ij}(u)$ denote the probability of making a transition from state i to state j when action u is taken from state i. Please refer Section 4.2 on Stationary Randomized Policies in Chapter 4, for information on notations. Λ denotes the set of stationary randomized policies (stochastic control kernels to be precise), whereas Υ denotes the set of stationary deterministic policies (control functions to be precise). Note that $|\Upsilon| = \prod_{i=1}^{n} |\mathcal{A}(i)|$, where $|\mathcal{A}|$ denotes the cardinality of the set $\mathcal{A}$. In this appendix we are not interested in the cost structure.

Any stationary deterministic policy μ will give rise to a Markov Chain (M.C.) with transition probability matrix P_μ, where $[P_\mu]_{i\,j} = p_{i\,j}(\mu(i))$. Similarly any stationary randomized policy $\delta \in \Lambda$ will give rise to a M.C. with transition probability matrix $P_\delta = [p_{ij}^{\delta}]$, where $p_{ij}^{\delta} = \sum_{a \in \mathcal{A}(i)} [\delta(i)]_a \; p_{ij}(a)$. Here $[\delta(i)]_a$ is the probability of taking action a from state i under the stationary randomized policy δ. A stationary fully randomized policy $\delta \in$ Interior(Λ) (see Section 4.2 on Stationary Randomized Policies in Chapter 4 for the definition of Interior(Λ)) is any stationary randomized

policy which assigns positive probability to each possible action from every state.

In this appendix we are concerned about the changes in the "reachability" structure as we go from stationary deterministic to stationary fully randomized policies.

B.1 Structure Of A General Stochastic Matrix

We borrow the terminology for the classification of the states from [43, pages 11–12]. Let $P = [p_{i\,j}],\ i,\, j\, =\, 1, \cdots, n;$ be any $n \times n$ stochastic matrix, which represents the transition probability matrix for some n state Markov Chain (M.C.). A sequence $(i, i_1, i_2, \cdots, i_{t-1}, j)$, for $t \geq 1$ (where $i_0 = i,\ i_t = j$), from the index set $\{1, 2, \ldots, n\}$ (states of the M.C.) is said to form a *chain* of length t between the ordered pair (i, j) if

$$p_{ii_1}\, p_{i_1 i_2} \cdots p_{i_{t-2} i_{t-1}}\, p_{i_{t-1} j} > 0$$

Such a chain for which $i = j$ is called a *cycle* of length t between i and itself. Without loss of generality we may impose the restriction that, for fixed $(i, j),\ i, j \neq i_1 \neq i_2 \neq \cdots \neq i_{t-1}$, to obtain a 'minimal' length chain or cycle, from a given one. Note that this does not preclude the possibility of i being the same as j.

B.1.1 Classification Of Indices For A Markov Chain

Let $i,\ j,\ k$ be arbitrary indices from the index set $\{1, 2, \ldots, n\}$ of the matrix P. For any positive integer m, let $p_{ij}^{(m)}$ denote the $(i, j)^{\text{th}}$ entry of P^m, the m^{th} power of P. We say that i *leads to* j, and write $i \longrightarrow j$, if there exists an integer $m \geq 1$ such

that $p_{ij}^{(m)} > 0$, or equivalently, if there is a chain between i and j. If i does not lead to j we write $i \not\longrightarrow j$. Clearly, if $i \longrightarrow j$ and $j \longrightarrow k$ then, from the rule of matrix multiplication, $i \longrightarrow k$. Note that for each i, there is some j (depending on i and the matrix P) such that $i \longrightarrow j$, since $\sum_{j=1}^{n} p_{ij} = 1 > 0$ for each i. We say that i and j *communicate* if $i \longrightarrow j$ and $j \longrightarrow i$, and denote it by $i \longleftrightarrow j$.

The indices of the stochastic matrix P, or equivalently the states of the M.C. can be classified and grouped as follows.

(a) If $i \longrightarrow j$ but $j \not\longrightarrow i$ for some j, then the index i is called *inessential*.

(b) Otherwise the index i is called *essential*. Thus if i is essential, $i \longrightarrow j$ implies $i \longleftrightarrow j$; and there is at least one j such that $i \longrightarrow j$.

(c) Hence it is clear that all essential indices can be subdivided into *essential classes* or *ergodic classes* in such a way, that all indices belonging to one class communicate, but cannot lead to an index outside the class. It can be proved that for a finite state M.C. there is at least one essential class [43, page 16].

(d) All inessential indices (if any) which communicate with some index, can be subdivided into *inessential classes* such that all indices in a class communicate. Note that any index which communicates with an index of an inessential class also belongs to that inessential class.

Classes of the type described in (c) and (d) are called *self-communicating* classes. Note that an index i belongs to some self-communicating class iff $i \longrightarrow i$ (or equivalently $i \longleftrightarrow i$).

(e) In addition there may be inessential indices which communicate with no index; these are defined as forming an *inessential class* by themselves (which, of course, if not self-communicating).

The inessential indices (or states) are also called *transient indices* (or *transient states*). Note that if a state i is transient and if j is such that $j \longrightarrow i$, then j too is transient. For any square non-negative matrix T (i.e. all the entries of T are non-negative real values) the corresponding *incidence matrix* $\tilde{T}$ replaces all the positive entries of T by ones. Note that the classification of indices (and hence grouping into classes) for the stochastic matrix (or equivalently the states of the M.C.) depends only on the location of the positive elements, and not on their magnitude, so any two stochastic matrices with the same incidence matrix will have the same index classification and grouping.

B.2 Rearrangement Of Index Classification, When We Move From Deterministic To Fully Randomized Policies

For any stationary deterministic policy μ of the MDP, we denote $i \xrightarrow{\mu} j$ iff i leads to j under the policy μ. We say $i \overset{\mu}{\not\longrightarrow} j$ iff i does not lead to j under the policy μ. Similarly we denote $i \xleftrightarrow{\mu} j$ iff $i \longleftrightarrow j$ under the policy μ. Similar notations hold for any stationary randomized policy δ.

Lemma B.1 *Let i, j be arbitrary indices from the index set $\{1, 2, \ldots, n\}$ (or states) and let δ be any stationary fully randomized policy. Then $i \xrightarrow{\delta} j$ iff $i \xrightarrow{\mu} j$ for at least one stationary deterministic policy μ.* □

Proof of Lemma B.1

It is easy to see that if $i \xrightarrow{\mu} j$ for some deterministic policy μ, then $i \xrightarrow{\delta} j$ from the definition of "leads to" and the fact that δ is a fully randomized policy.

Now the only if part can be proved as follows. Let i, j be such that $i \xrightarrow{\delta} j$. Hence we can find a chain $(i, i_1, i_2, \cdots, i_{t-1}, j)$ of length t between the ordered pair (i, j) for the M.C. with transition probability matrix P^{δ}. Without loss of generality we may assume this to be a minimal length chain. Note that $0 < t \leq n$. With the notation that $i_0 = i$ and $i_t = j$, we have that $p^{\delta}_{i_k\, i_{k+1}} > 0$ for $0 \leq k \leq t-1$. Hence there exists actions $a_k \in \mathcal{A}(i_k)$ for $0 \leq k \leq t-1$ such that $p_{i_k\, i_{k+1}}(a_k) > 0$ for $0 \leq k \leq t-1$. Pick any stationary deterministic policy μ such that $\mu(i_k) = a_k$ for each $0 \leq k \leq t-1$. Then $i \xrightarrow{\mu} j$.

□

Lemma B.2 *An index $i \in \{1, 2, \ldots, n\}$ belongs to some self-communicating class for a stationary fully randomized policy δ iff i belongs to a self-communicating class for some stationary deterministic policy μ.*

□

Proof of Lemma B.2

Apply Lemma B.1 with $i = j$ and use the fact that for any Markov Chain, i belongs to one of its self communicating classes iff $i \longrightarrow i$ for this Markov Chain.

□

Let m_μ represent the number of distinct ergodic classes for the stationary deterministic policy μ. Similarly, let m_δ denote the number of distinct ergodic classes for the stationary randomized policy δ. Let $C_1^\mu, C_2^\mu, \ldots, C_{m_\mu}^\mu$ be the ergodic classes for the deterministic policy μ. Similarly for any stationary randomized policy δ, let $C_1^\delta, C_2^\delta, \ldots, C_{m_\delta}^\delta$ be its ergodic classes. We have the following results.

Theorem B.1 *Let μ be a stationary deterministic policy and let δ be a stationary fully randomized policy. Then*

1. *Given any C_i^δ and any μ, we can find a j such that $C_j^\mu \subseteq C_i^\delta$.*
2. $$m_\delta = \min_{\mu \in \Upsilon} m_\mu$$
3. $$\bigcup_{i=1}^{m_\delta} C_i^\delta = \bigcup_{\{\mu : m_\mu = m_\delta\}} \bigcup_{i=1}^{m_\mu} C_i^\mu$$

□

The proof of Theorem B.1 is given later. The above theorem implies that for any stationary deterministic policy μ with $m_\mu = m_\delta$, precisely one of its ergodic classes will be a subset of each ergodic class of any stationary fully randomized policy δ. Note that if a state i belongs to an ergodic class of δ then a state j is in the same ergodic class of δ iff $i \xrightarrow{\delta} j$. Please refer Lemma B.1.

Corollary B.1 *If a state i is transient for every stationary deterministic policy μ, then i is transient for any stationary fully randomized policy δ. Equivalently if a state i belongs to some ergodic class for a stationary fully randomized policy, then i belongs to an ergodic class for some stationary deterministic policy.*

□

Proof of Corollary B.1

Refer the proof of Theorem B.1.

□

Corollary B.2 *If a state i belongs to some ergodic class for every stationary deterministic policy μ, then i belongs to some ergodic class for any stationary fully randomized policy δ. Equivalently if a state i is transient for a stationary fully randomized policy, then i is transient for some stationary deterministic policy.*

□

Proof of Corollary B.2

Refer the proof of Theorem B.1.

□

Let $B \subseteq \mathcal{S}$, be nonempty. For any stationary fully randomized policy δ and any stationary deterministic policy μ we denote

$$B_\delta = \{i \in \mathcal{S} \,|\, i \xrightarrow{\delta} j, \text{ for some } j \in B\}$$

and

$$B_\mu = \{i \in \mathcal{S} \,|\, i \xrightarrow{\mu} j, \text{ for some } j \in B\}$$

Note that B_δ (respectively B_μ) is constituted of precisely those states which *lead to* some state in B under policy δ (respectively μ). Observe that these sets can be empty. In the following algorithm and discussions, we say that a state $i \in \mathcal{S}$ is *marked* if we assign a particular action $a \in \mathcal{A}(i)$ to the state i.

Before we prove Theorem B.1, we prove the following lemma.

Lemma B.3 *Let $B \subseteq \mathcal{S}$ be nonempty and δ be any stationary fully randomized policy. Then there exists a marking of the states in $B_\delta \setminus B$ such that any stationary deterministic policy μ which agrees on the actions taken from the set $B_\delta \setminus B$ with the above mentioned marking, has $B_\delta \setminus B = B_\mu \setminus B$.*

□

Proof of Lemma B.3

Using Lemma B.1 we have that if $\tilde{\mu}$ is any stationary deterministic policy, then $B_{\tilde{\mu}} \subseteq B_\delta$. Hence $B_\mu \setminus B \subseteq B_\delta \setminus B$. Hence we only need to prove that there exists a marking of the states in $B_\delta \setminus B$ such that any stationary deterministic policy μ which agrees on the actions taken from the set $B_\delta \setminus B$ with the above mentioned marking has $B_\delta \setminus B \subseteq B_\mu \setminus B$.

If $B_\delta \setminus B$ is empty, we have nothing to prove. Suppose $B_\delta \setminus B$ is nonempty. We assume that the states in $B_\delta \setminus B$ are not marked initially. Then we use the following algorithm to mark the states in $B_\delta \setminus B$ and the result follows.

□

A systematic algorithm which does not assume the apriori knowledge of the set $B_\delta \setminus B$ is given below.

Algorithm B.1

- *Initialize* $C = \mathcal{S} \setminus B$, $\tilde{C} = \emptyset$, $\mathcal{S}_0 = B$, $k = 0$.
- *While* C *is nonempty and there exists* $i \in C$ *such that* $p_{ij}^{\delta} > 0$ *with* $j \in \mathcal{S}_k$, *do the following.*

  (a) *Set* $\mathcal{S}_{k+1} = \emptyset$.

  (b) *For all* $i \in C$ *do*

  If i *is such that* $p_{ij}^{\delta} > 0$ *for some* $j \in \mathcal{S}_k$

  1. *Pick* $a \in \mathcal{A}(i)$ *such that* $p_{ij}(a) > 0$. *It is easy to see that such an* a *exists.*
  2. *Mark state* i *with the corresponding action* $a \in \mathcal{A}(i)$ *obtained from the previous step. Remove state* i *from the set* C. *Add state* i *to the set* $\mathcal{S}_{k+1}$.

  (c) *Set* $\tilde{C} = \tilde{C} \cup \mathcal{S}_{k+1}$.

  (d) *Set* $k = k + 1$.

□

The while loop will iterate at most $|\mathcal{S} \setminus B|$ times. It is clear from Algorithm B.1, that at the end of each iteration of the while loop, all the states which have already been marked until that iteration (i.e. $\tilde{C}$), belong to the set $B_{\tilde{\mu}} \setminus B$ for any stationary

deterministic policy $\tilde{\mu}$ which agree on the already marked states $\tilde{C}$. Note that in Algorithm B.1, the while loop condition is true as long as $C \cap (B_\delta \setminus B) \neq \emptyset$. When the algorithm terminates, $\tilde{C}$ will be equal to $B_\delta \setminus B$.

Corollary B.3 *Let δ be any stationary fully randomized policy. Suppose that the M.C. with transition probability matrix P_δ is unichain, i.e. it has only one ergodic class. Let B be any nonempty subset of this ergodic class. Suppose all the states of this set B are marked (i.e. for each state i in the set B, we assign a particular action a_i from the set $\mathcal{A}(i)$), and the corresponding marking on the set B gives rise to a M.C. over the subset B, i.e. $p_{ij}(a_i) = 0,\ \forall j \notin B, i \in B$. Then there exists a deterministic policy μ, which agrees with the afore mentioned markings on the set B, and having the property that $B_\mu = \mathcal{S}$. Also the ergodic classes for the M.C. with transition probability matrix P_μ are the same as the ergodic classes for the M.C. restricted to the subset B.*

□

Proof of Corollary B.3

First note that $B_\delta = \mathcal{S}$, since the M.C. corresponding to the transition probability matrix P_δ is unichain and B is a subset of this unique ergodic class for this Markov Chain. By Lemma B.3, we can find a stationary deterministic policy μ, which agrees with the markings (mentioned in the statememt of Corollary B.3) on the set B, and having the property that $B_\mu \setminus B = \mathcal{S} \setminus B$. By the choice of the markings on B, we have $B_\mu \cap B = B$. Hence $B_\mu = \mathcal{S}$. This along with the fact that, the M.C. restricted to the set B (for the deterministic policy μ) is a M.C. tells

us that B is an absorbing set for the policy μ (i.e. $p_{ij}(\mu(i)) = 0,\ \forall j \notin B, i \in B$). Note that $i \in \mathcal{S} \setminus B$ imply that i is an inessential index for the M.C. with transition probability matrix P_μ. Hence the ergodic classes for the M.C. with transition probability matrix P_μ are the same as the ergodic classes for the M.C. restricted to the subset B.

□

Proof of Theorem B.1

First of all we prove that given any C_i^δ and any μ then we can find a j such that $C_j^\mu \subseteq C_i^\delta$. Fix any deterministic policy μ and any one of the ergodic classes C_i^δ. Let $k \in C_i^\delta$. Then $k \overset{\mu}{\not\longrightarrow} l$ for any $l \notin C_i^\delta$, since C_i^δ is an ergodic class for the fully randomized policy δ (refer Lemma B.1). But there exists $j \in \{1, \ldots, m_\mu\}$ such that $k \in (C_j^\mu)_\mu$ where

$$(C_j^\mu)_\mu = \{l \in \mathcal{S} \mid l \overset{\mu}{\longrightarrow} \tilde{l},\ \text{for some } \tilde{l} \in C_j^\mu\}$$

and this in turn implies $C_j^\mu \subseteq C_i^\delta$ (again by referring to Lemma B.1). This also implies that $m_\mu \geq m_\delta$. From the above arguments, it is clear that for any deterministic policy $\tilde{\mu}$ with $m_{\tilde{\mu}} = m_\delta$ (if at all it exists), precisely one each of its ergodic classes will be a subset of each C_k^δ for $k \in \{1, \ldots, m_\delta\}$. Hence

$$\bigcup_{\{\mu\,:\,m_\mu = m_\delta\}} \bigcup_{i=1}^{m_\mu} C_i^\mu \ \subseteq \ \bigcup_{k=1}^{m_\delta} C_k^\delta$$

Now to prove the last statement of the theorem. Pick arbitrary $i_1, \ldots, i_{m_\delta}$ such that $i_k \in C_k^\delta$ for $k \in \{1, \ldots, m_\delta\}$. Let $B \equiv \{i_1, \ldots, i_{m_\delta}\}$. Notice that $B_\delta = \mathcal{S}$. Then by Lemma B.3 there exists marking of the states in $\mathcal{S} \setminus B$ (i.e. a particular

assignment of action to each state in $\mathcal{S} \setminus B$, say action a_l to state $l \in \mathcal{S} \setminus B$) such that any deterministic policy μ with $\mu(l) = a_l,\ \forall l \in \mathcal{S} \setminus B$, has $\mathcal{S} \setminus B = B_\mu \setminus B$. Now for $l \in B$, pick any $a_l \in \mathcal{A}(l)$ and assign $\mu(l) = a_l$. Note that under this choice of μ, $l \xrightarrow{\mu} i_k$ for each $l \in C_k^\delta$. Also $l \in C_k^\delta$ implies $l \overset{\mu}{\not\longrightarrow} \tilde{l}$, for all $\tilde{l} \notin C_k^\delta$. This also implies that there is exactly one ergodic class of this policy μ in each of the sets C_k^δ and i_k is an element of this ergodic class. Since $\mathcal{S} \setminus (\bigcup_{k=1}^{m_\delta} C_k^\delta) \subseteq B_\mu \setminus B$, the states in $\mathcal{S} \setminus (\bigcup_{k=1}^{m_\delta} C_k^\delta)$ are transient under the deterministic policy μ (by virtue of the choice of the set B). This also implies that $m_\mu = m_\delta$. Hence we have $m_\delta = \min_{\tilde{\mu} \in \Gamma} m_{\tilde{\mu}}$.

Since i_k could have been any state in C_k^δ in the choice of the set B, we have

$$\bigcup_{k=1}^{m_\delta} C_k^\delta = \bigcup_{\{\mu\,:\,m_\mu = m_\delta\}} \bigcup_{i=1}^{m_\mu} C_i^\mu$$

Also this imples Corollaries B.1 and B.2.

□

Appendix C

Error Bounds For Markov Decision Processes

In this appendix we discuss some results related to the error bounds for Markov Decision Processes (MDPs).

In Section C.1 we discuss a general contraction mapping theorem [12, 37], approximate value iteration and some generic error bounds for contraction mappings. In Section C.2 we deal with the Stochastic Shortest Path problem (SSP) model and discuss absorption probability issues of SSPs and explore the average number of stages needed to reach the terminal state. Section C.3 discusses issues related to properness and acyclicity of policies in SSP. In Section C.4 we discuss the contraction properties of SSP dynamic programming operator along with various error bounds for SSP. Section C.5 deals with the equivalent SSP problem for discounted cost MDP. Various error bounds for discounted cost problem are dealt with. In Section C.6 error bounds for average cost problem are dealt with.

C.1 Contraction Mappings

Let $\mathcal{V}$ be a Banach Space [41] that is a normed linear space which is complete under the norm $\| \cdot \|$.

Let $H : \mathcal{V} \rightarrow \mathcal{V}$ be a mapping such that $\| HV - HV' \| \leq K \| V - V' \|$, $\forall\ V, V' \in \mathcal{V}$; where $0 \leq K < \infty$. Then H is a uniformly continuous mapping. Here

HV is the mapping H applied to $V \in \mathcal{V}$. If $\| HV - HV' \| \leq \| V - V' \| \quad \forall\, V, V' \in \mathcal{V}$ then H is called a nonexpansion mapping.

A mapping $H : \mathcal{V} \to \mathcal{V}$ is said to be a contraction mapping with modulus of contraction α, if there exists a scalar α with $0 \leq \alpha < 1$, such that

$$\| HV - HV' \| \leq \alpha \| V - V' \| \qquad \forall\, V, V' \in \mathcal{V}$$

$H : \mathcal{V} \to \mathcal{V}$ is said to be an m-stage contraction mapping if there exists a positive integer m and some scalar α, with $0 \leq \alpha < 1$ such that

$$\| H^m V - H^m V' \| \leq \alpha \| V - V' \| \qquad \forall\, V, V' \in \mathcal{V}$$

Here H^m denotes the composition of H with itself m times; i.e. $H^{k+1}V = HH^kV$, $\forall\ V \in \mathcal{V},\ k = 0, 1, 2, \cdots$. H^0 is the identity mapping, i.e. $H^0V = V$, $\forall\ V \in \mathcal{V}$. Again α is called the modulus of contraction. The modulus of contraction is also called the contraction coefficient. Note that if H is a contraction mapping, then H is uniformly continuous. However H being an m-stage contraction does not necessarily imply that H is continuous. As in the following example, H may be discontinuous every where.

Example C.1 *Let $H : \mathbf{R} \to \mathbf{R}$ be such that*

$$Hx = 0 \qquad x \text{ rational}$$

$$Hx = 1 \qquad x \text{ irrational}$$

Let $\| x \| = |x|$, $\forall\ x \in \mathbf{R}$. H is discontinuous everywhere and $H^2x = 0$, $\forall x \in \mathbf{R}$. Also $H^kx = 0$, $\forall k \geq 2, x \in \mathbf{R}$. Hence

$$\| H^k x - H^k y \| \leq 0 \cdot \| x - y \| \qquad \forall k \geq 2,\ x, y \in \mathbf{R}$$

Hence the modulus of contraction may be chosen to be 0 *and* H *is a two stage contraction.*

□

C.1.1 Contraction Mapping Theorem

Proposition C.1 (Contraction Mapping Fixed Point Theorem) *Let* $H : \mathcal{V} \to \mathcal{V}$ *be a contraction mapping (i.e. a one stage contraction mapping) or an* m*-stage contraction mapping for some positive integer* m*. Let* α *where* $0 \leq \alpha < 1$ *be the contraction coefficient. Then there exists a unique fixed point* $V^* \in \mathcal{V}$ *such that*

1. $$HV^* = V^*$$

2. *Furthermore if* V *is any element in* $\mathcal{V}$ *and* H^k *is the composition of* H *with itself* k *times for* $k \geq 0$*, then*

$$\lim_{k\to\infty} \| H^k V - V^* \| = 0$$

Also

$$\begin{aligned} \| H^{km+l} V - V^* \| &\leq \alpha^k \| H^l V - V^* \| \\ &\leq \alpha^k \left(\max_{\tilde{l} \in \{0,1,\ldots,m-1\}} \| H^{\tilde{l}} V - V^* \| \right) \end{aligned}$$

for $l = 0, 1, \ldots, m-1$ *and* $k \geq 0$*.*

□

We provide a proof of the result below. Note that we do not assume H to be continuous as in the proof given in [37]. See [12] for an alternate proof.

Proof of Proposition C.1

Let H be an m-stage contraction mapping for some integer $m > 0$ and let the contraction coefficient be α where $0 \leq \alpha < 1$. Let $V \in \mathcal{V}$. Define

$$V_k = H^k V \qquad \text{for integer } k \geq 0$$

with $V_0 = H^0 V = V$. Notice that for all integers $k \geq 0$, $l \geq 0$

$$\| H^{ml} V_{k+1} - H^{ml} V_k \| \leq \alpha^l \| V_{k+1} - V_k \|$$

Let

$$\bar{K} = \max_{k \in \{1,2,\ldots,m\}} \| V_k - V_{k-1} \|$$

Note that for $0 \leq k < m$ and $l \geq 0$

$$\| V_{ml+k+1} - V_{ml+k} \| \leq \alpha^l \bar{K}$$

Hence for $l \geq 0$

$$\begin{aligned} \sum_{k=0}^{\infty} \| V_{ml+k+1} - V_{ml+k} \| &\leq m\bar{K} \sum_{i=l}^{\infty} \alpha^i \\ &= m\bar{K} \frac{\alpha^l}{1-\alpha} \end{aligned}$$

Hence $\{V_k\}$ is a Cauchy sequence. Thus there exists $V^* \in \mathcal{V}$ such that $V_k \overset{k \to \infty}{\longrightarrow} V^*$; i.e. $\lim_{k \to \infty} \| V_k - V^* \| = 0$. Let $\bar{V} \in \mathcal{V}$ and

$$\bar{V}_k = H^k \bar{V} \qquad \text{for integer } k \geq 0$$

with $\bar{V}_0 = H^0\bar{V} = \bar{V}$. Note that $\{\bar{V}_k\}$ is a Cauchy sequence and hence converges to some $\bar{V}^* \in \mathcal{V}$. Consider the subsequence $\{V_{ml}\}$ and $\{\bar{V}_{ml}\}$ where $l \in \mathbf{N}_0$ (Note that $\mathbf{N}_0$ is the set of non-negative integers). Now $V_{ml} \overset{l\to\infty}{\longrightarrow} V^*$ and $\bar{V}_{ml} \overset{l\to\infty}{\longrightarrow} \bar{V}^*$. But

$$\| V_{ml} - \bar{V}_{ml} \| \leq \alpha^l \| V - \bar{V} \|$$

Thus $\lim_{l\to\infty} \| V_{ml} - \bar{V}_{ml} \| = 0$, implying $V^* = \bar{V}^*$. Thus irrespective of the starting element $V \in \mathcal{V}$, the sequence $\{H^kV\}$, $k \in \mathbf{N}_0$ converges to the unique vector V^*.

Since H^m is a contraction mapping with contraction coefficient α, we have that H^m is continuous. Now $V_{mk} \overset{k\to\infty}{\longrightarrow} V^*$. Also $H^mV_{mk} = V_{m(k+1)}$. Hence $H^mV^* = V^*$. We will prove that for $m > 1$, $HV^* = V^*$. Suppose not, i.e. $HV^* = \tilde{V}$ and $V^* \neq \tilde{V}$. Thus $\| V^* - \tilde{V} \| > 0$. Let us start the iteration with $V_0 = V^*$ and $V_k = H^kV_0$ for $k \geq 0$. Now $V_{ml} = H^{ml}V_0 = H^{ml}V^* = V^*$ for $l \geq 0$. Now $V_{ml+1} = H^{ml+1}V_0 = H^{ml+1}V^* = HV^* = \tilde{V}$. Now the sequence $\{V_k = H^kV^*\}$ converges to V^* by our earlier discussion. Hence any subsequence of $\{H^kV^*\}$, $k \in \mathbf{N}_0$ should converge to V^*. But the subsequence $\{V_{ml+1} = H^{ml+1}V^*\}$, $l \in \mathbf{N}_0$ converges to $\tilde{V} \neq V^*$. Hence we have a contradiction. Thus $HV^* = V^*$.

Uniqueness of the fixed point follows immediately. Let V' and V^* be fixed points of H. Then $H^kV' = V'$ for $k \geq 0$ and likewise $H^kV^* = V^*$ for $k \geq 0$. Now

$$\| H^{ml}V' - H^{ml}V^* \| = \| V' - V^* \|$$

for all $l \geq 0$. But since H^m is a contraction mapping

$$\| H^{ml}V' - H^{ml}V^* \| \leq \alpha^l \| V' - V^* \|$$

for all $l \geq 0$. Here $0 \leq \alpha < 1$. Thus

$$\| V' - V^* \| \leq \alpha^l \| V' - V^* \|$$

forall $l \geq 0$, implying $\| V' - V^* \| = 0$; i.e. $V' = V^*$. Part 2 of the proposition follows immediately from the definition of m-stage contraction mapping.

□

Note that H need not be continuous at the fixed point V^* as was shown in Example C.1.

Proposition C.2 *Let $H : \mathcal{V} \to \mathcal{V}$ be an m-stage contraction mapping with $m > 0$, where m is an integer. Let α with $0 \leq \alpha < 1$ be the contraction coefficient. Assume further that H is a non-expansion. Suppose that for some $V \in \mathcal{V}$, $\| HV - V \| \leq \epsilon$. Then $\| V^* - V \| \leq \frac{m\epsilon}{1-\alpha}$ where V^* is the unique fixed point of H.*

□

Proof of Proposition C.2

Now $\| HV - HV' \| \leq \| V - V' \|$ and $\| H^m V - H^m V' \| \leq \alpha \| V - V' \|$ for all $V, V' \in \mathcal{V}$. Let $V_k = H^k V$ for $k \geq 0$, with $V_0 = H^0 V = V$. Note that $\| V_k - V_{k-1} \| \leq \epsilon$ for $k \in \{1, 2, \ldots, m\}$. Also

$$\| V_{ml+k+1} - V_{ml+k} \| \leq \alpha^l \| V_{k+1} - V_k \| \qquad \forall\, l \geq 0, k \geq 0$$

Hence for $k \in \{0, 1, \ldots, m-1\}$ and $l \geq 0$

$$\| V_{ml+k+1} - V_{ml+k} \| \leq \alpha^l \epsilon$$

Thus

$$\begin{aligned}\sum_{j=0}^{\infty} \parallel V_{j+1} - V_j \parallel &= \sum_{l=0}^{\infty}\sum_{k=0}^{m-1} \parallel V_{ml+k+1} - V_{ml+k} \parallel \\ &\leq \sum_{l=0}^{\infty} \alpha^l m\epsilon \\ &= \frac{m\epsilon}{1-\alpha}\end{aligned}$$

Since $V_k \stackrel{k\to\infty}{\longrightarrow} V^*$, given any scalar $\varepsilon > 0$, there exists $k' > 0$ such that $\parallel V_k - V^* \parallel \leq \varepsilon$ for all $k \geq k'$. Thus

$$\begin{aligned}\parallel V - V^* \parallel &= \parallel V_0 - V^* \parallel \\ &\leq \parallel V_0 - V_{k'} \parallel + \parallel V_{k'} - V^* \parallel \\ &\leq \sum_{k=0}^{k'-1} \parallel V_{k+1} - V_k \parallel + \varepsilon \\ &\leq \sum_{k=0}^{\infty} \parallel V_{k+1} - V_k \parallel + \varepsilon \\ &\leq \frac{m\epsilon}{1-\alpha} + \varepsilon\end{aligned}$$

Since this is true for any $\varepsilon > 0$, we have

$$\parallel V - V^* \parallel \leq \frac{m\epsilon}{1-\alpha}$$

□

Example C.2 *Consider the earlier example where* $\mathcal{V} = \mathbf{R}$ *and* $H : \mathbf{R} \to \mathbf{R}$ *is such that*

$$\begin{aligned}Hx &= 0 \quad x \text{ rational} \\ Hx &= 1 \quad x \text{ irrational}\end{aligned}$$

Let $\| x \| = |x|$, $\forall\, x \in \mathbf{R}$. *Now* H *is a* 2*-stage contraction mapping with contraction coefficient* 0. *However* H *is not a non-expansion. Let* $x_k \stackrel{k\to\infty}{\longrightarrow} 1$, x_k *irrational. Hence* $Hx_k = 1$ *for all* $k \geq 0$ *and* $\| Hx_k - x_k \| \stackrel{k\to\infty}{\longrightarrow} 0$. *But the unique fixed point of* H *is* $x^* = 0$ *and* $\lim_{k\to\infty} \| x_k - x^* \| = \lim_{k\to\infty} \| x_k \| = 1$.

Thus for general m-stage contraction mappings (for $m > 1$), it is not true that if $\| HV - V \|$ is "small", then $\| V - V^* \|$ is "small".

□

C.1.2 Approximate Value Iteration

Lemma C.1 (Approximate Value Iteration) *Let* $\mathcal{V}$ *be a Banach space, i.e. a normed linear space which is complete under a norm* $\| \cdot \|$. *Let* $H : \mathcal{V} \to \mathcal{V}$ *be a non-expansive mapping which is an* m*-stage contraction for some integer* $m > 0$. *That is*

$$\| HV - HV' \| \leq \| V - V' \| \qquad \forall\, V, V' \in \mathcal{V}$$

and

$$\| H^m V - H^m V' \| \leq \alpha \| V - V' \| \qquad \forall\, V, V' \in \mathcal{V}$$

Here α *is the contraction coefficient and* $0 \leq \alpha < 1$.

Consider the approximate value iteration method that generates a sequence $\{V_k\}$, *with* $V_k \in \mathcal{V}$ *satisfying*

$$\| V_{k+1} - HV_k \| \leq \epsilon$$

for $k \geq 0$ *and some scalar* $\epsilon \geq 0$, *starting from an arbitrary* $V_0 \in \mathcal{V}$. *Let* V^* *be the*

unique fixed point of H. Then

$$\limsup_{k\to\infty} \| V_k - V^* \| \leq \frac{m\epsilon}{1-\alpha}$$

□

Proof of Lemma C.1

Note that $\| V_0 \| < \infty$ and that if $\epsilon = 0$ we have value iteration.

Let l be a non-negative integer. Then $\| V_{l+1} - HV_l \| \leq \epsilon$. Hence $\| HV_{l+1} - H^2V_l \| \leq \epsilon$. Now $\| V_{l+2} - HV_{l+1} \| \leq \epsilon$. Hence

$$\begin{aligned} \| V_{l+2} - H^2V_l \| &\leq \| V_{l+2} - HV_{l+1} \| + \| HV_{l+1} - H^2V_l \| \\ &\leq 2\epsilon \end{aligned}$$

Now $\| HV_{l+2} - H^3V_l \| \leq 2\epsilon$ and $\| V_{l+3} - HV_{l+2} \| \leq \epsilon$. Hence

$$\begin{aligned} \| V_{l+3} - H^3V_l \| &\leq \| V_{l+3} - HV_{l+2} \| + \| HV_{l+2} - H^3V_l \| \\ &\leq 3\epsilon \end{aligned}$$

Continuing similarly

$$\| V_{l+m} - H^mV_l \| \leq m\epsilon \tag{C.1}$$

Now

$$\| H^mV_{l+m} - H^{2m}V_l \| \leq \alpha\, m\epsilon$$

since H is an m-stage contraction. Now by the inequality C.1,

$\| V_{2m+l} - H^mV_{m+l} \| \leq m\epsilon$. Hence

$$\| V_{2m+l} - H^{2m}V_l \| \leq \| V_{2m+l} - H^mV_{m+l} \| + \| H^mV_{m+l} - H^{2m}V_l \|$$

$$\leq \quad m\epsilon + \alpha m\epsilon$$

$$= \quad (1+\alpha)m\epsilon$$

Now $\| H^m V_{2m+l} - H^{3m} V_l \| \leq (\alpha + \alpha^2) m\epsilon$. Also by inequality C.1,

$\| V_{3m+l} - H^m V_{2m+l} \| \leq m\epsilon$. Hence

$$\begin{aligned} \| V_{3m+l} - H^{3m} V_l \| &\leq \| V_{3m+l} - H^m V_{2m+l} \| + \| H^m V_{2m+l} - H^{3m} V_l \| \\ &\leq m\epsilon + (\alpha + \alpha^2) m\epsilon \\ &= (1 + \alpha + \alpha^2) m\epsilon \end{aligned}$$

Continuing similarly or by an induction argument, it is true that for any integer $k \geq 1$ and $l \geq 0$

$$\begin{aligned} \| V_{km+l} - H^{km} V_l \| &\leq (1 + \alpha + \alpha^2 + \ldots + \alpha^{k-1}) m\epsilon \\ &\leq \frac{m\epsilon}{1-\alpha} \end{aligned}$$

Hence

$$\limsup_{k \to \infty} \| V_{km+l} - H^{km} V_l \| \leq \frac{m\epsilon}{1-\alpha}$$

Let V^* be the unique fixed point of H. Now

$$\begin{aligned} \| V_{km+l} - V^* \| &\leq \| V_{km+l} - H^{km} V_l \| + \| H^{km} V_l - V^* \| \\ &\leq \frac{m\epsilon}{1-\alpha} + \| H^{km} V_l - V^* \| \end{aligned}$$

Since

$$\lim_{k \to \infty} \| H^{km} V_l - V^* \| = 0$$

by the contraction mapping fixed point Theorem C.1, we have

$$\limsup_{k \to \infty} \| V_{km+l} - V^* \| \quad \leq \quad \frac{m\epsilon}{1-\alpha} \tag{C.2}$$

Since the inequality C.2 is true for $l = 0, 1, 2, \ldots, m-1$ we have

$$\limsup_{k \to \infty} \| V_k - V^* \| \leq \frac{m\epsilon}{1-\alpha}$$

□

C.1.3 Contraction Mapping Generic Error Bounds

Lemma C.2 *Let H and $\tilde{H}$ both be contraction mappings (one stage) with contraction coefficient, α (where $0 \leq \alpha < 1$) under some norm $\| \cdot \|$ on a Banach space $\mathcal{V}$. Let V^* be the unique fixed point of H. Suppose $V \in \mathcal{V}$ be such that $\| V - V^* \| \leq \epsilon$ and $\| HV - \tilde{H}V \| \leq \varepsilon$ where scalars $\epsilon \geq 0, \varepsilon \geq 0$. Then*

$$\| \tilde{V} - V^* \| \leq \frac{2\alpha\epsilon + \varepsilon}{1-\alpha}$$

where $\tilde{V}$ is the unique fixed point of $\tilde{H}$.

□

Proof of Lemma C.2

Note that both H and $\tilde{H}$ have the same contraction coefficient α. Since $HV^* = V^*$

$$\begin{aligned} \| \tilde{H}V - V \| &\leq \| \tilde{H}V - HV \| + \| HV - HV^* \| + \| V^* - V \| \\ &\leq \varepsilon + \alpha \| V - V^* \| + \| V - V^* \| \\ &= \varepsilon + (1+\alpha) \| V - V^* \| \end{aligned}$$

Hence

$$\| \tilde{H}^2 V - \tilde{H}V \| \leq \alpha(\varepsilon + (1+\alpha) \| V - V^* \|)$$

By Proposition C.2 given earlier we have

$$\| \tilde{H}V - \tilde{V} \| \leq \frac{\alpha(\varepsilon + (1+\alpha) \| V - V^* \|)}{1-\alpha}$$

Now $HV^* = V^*$. Hence

$$\begin{aligned} \| V^* - \tilde{V} \| &\leq \| HV^* - HV \| + \| HV - \tilde{H}V \| + \| \tilde{H}V - \tilde{V} \| \\ &\leq \alpha \| V - V^* \| + \varepsilon + \frac{\alpha(\varepsilon + (1+\alpha) \| V - V^* \|)}{1-\alpha} \\ &= \frac{2\alpha \| V - V^* \| + \varepsilon}{1-\alpha} \\ &= \frac{2\alpha\epsilon + \varepsilon}{1-\alpha} \end{aligned}$$

□

We have the following extension for m-stage contraction mappings.

Lemma C.3 *Let H be a non-expansive mapping on a Banach space $\mathcal{V}$ under the norm $\| \cdot \|$. Let V^* be a fixed point of H. Let $\tilde{H}$ be a non-expansive mapping which is an m-stage ($m > 1$) contraction mapping with contraction coefficient α ($0 \leq \alpha < 1$), defined on the Banach space $\mathcal{V}$ under the same norm $\| \cdot \|$. Let $V \in \mathcal{V}$ be such that $\| V - V^* \| \leq \epsilon$ and $\| HV - \tilde{H}V \| \leq \varepsilon$ where scalars $\epsilon \geq 0, \varepsilon \geq 0$. Then*

$$\| V^* - \tilde{V} \| \leq \frac{(2(m-1) + (1+\alpha))\epsilon + m\varepsilon}{1-\alpha}$$

where $\tilde{V}$ is the unique fixed point of $\tilde{H}$.

□

Proof of Lemma C.3

Since $HV^* = V^*$

$$\begin{aligned} \| \tilde{H}V - V \| &\leq \| \tilde{H}V - HV \| + \| HV - HV^* \| + \| V^* - V \| \\ &\leq \varepsilon + \| V - V^* \| + \| V - V^* \| \\ &= \varepsilon + 2 \| V - V^* \| \end{aligned}$$

Also

$$\| \tilde{H}^l V - \tilde{H}^{l-1} V \| \leq \varepsilon + 2 \| V - V^* \| \qquad \text{for} l = 1, 2, \ldots, m$$

For $l = 1, 2, \ldots, m$ and $k = 0, 1, 2, \ldots$ we have

$$\| \tilde{H}^{km+l} V - \tilde{H}^{km+l-1} V \| \leq \alpha^k (\varepsilon + 2 \| V - V^* \|)$$

Since $\lim_{k \to \infty} \tilde{H}^k V = \tilde{V}$

$$\begin{aligned} \| \tilde{H}^m V - \tilde{V} \| &\leq \sum_{k=1}^{\infty} \sum_{l=1}^{m} \| \tilde{H}^{km+l} V - \tilde{H}^{km+l-1} V \| \\ &\leq \frac{m\alpha(\varepsilon + 2 \| V - V^* \|)}{1 - \alpha} \end{aligned}$$

Hence

$$\begin{aligned} \| \tilde{H}V - \tilde{V} \| &\leq \sum_{l=1}^{m-1} \| \tilde{H}^{l+1} V - \tilde{H}^l V \| + \| \tilde{H}^m V - \tilde{V} \| \\ &\leq (m-1)(\varepsilon + 2 \| V - V^*) + \frac{m\alpha(\varepsilon + 2 \| V - V^* \|)}{1 - \alpha} \end{aligned}$$

Since $HV^* = V^*$, we have

$$\begin{aligned} \| V^* - \tilde{V} \| &\leq \| HV^* - HV \| + \| HV - \tilde{H}V \| + \| \tilde{H}V - \tilde{V} \| \\ &\leq \| V - V^* \| + \varepsilon + (m-1)(\varepsilon + 2 \| V - V^* \|) + \frac{m\alpha(\varepsilon + 2 \| V - V^* \|)}{1 - \alpha} \\ &= \frac{(2(m-1) + (1+\alpha)) \| V - V^* \| + m\varepsilon}{1 - \alpha} \\ &= \frac{(2(m-1) + (1+\alpha))\epsilon + m\varepsilon}{1 - \alpha} \end{aligned}$$

□

C.2 Stochastic Shortest Path MDPs Revisited

Consider a homogeneous discrete time Stochastic Shortest Path (SSP) problem. For detailed notations on MDP see Chapter 1. We briefly state the notations for SSP MDPs here.

The finite state space is $\mathcal{S} = \{0, 1, 2, \ldots, n\}$ with state '0' being the termination state or the absorption state. $\mathcal{A}(i), i \in \mathcal{S}$ denotes the finite set of possible actions from state $i \in \mathcal{S}$, with $\mathcal{A}(i) = \{1, 2, \ldots, |\mathcal{A}(i)|\}$, where $|\mathcal{A}(i)|$ denotes the cardinality of the set $\mathcal{A}(i)$. Let $\mathcal{A} = \bigcup_{i \in \mathcal{S}} \mathcal{A}(i)$ denote the action space. The transition probabilities may be conveniently denoted by $p_{ij}(u) = \Pr\{s_{t+1} = j \mid s_t = i, u_t = u\}$, where 'Pr' denotes probability, $s_t \in \mathcal{S}$ denotes the state at time t, u_t denotes the action taken at time t from state s_t (here $u_t \in \mathcal{A}(s_t)$). Let g_t denote the immediate cost incurred at time t when action u_t is taken from state s_t and the system moves to state s_{t+1} at time $t + 1$. For $i, j \in \mathcal{S}$ and $u \in \mathcal{A}(i)$, let $g(i, u, j) \equiv \mathrm{E}[g_t \mid s_t = i, u_t = u, s_{t+1} = j]$, where 'E' denotes expectation. The expected immediate cost of taking action u from state i for $i \in \mathcal{S}, u \in \mathcal{A}(i)$ is

$$\begin{aligned} g(i, u) &\equiv \mathrm{E}[g_t \mid s_t = i, u_t = u] \\ &= \sum_{j=0}^{n} p_{ij}(u) g(i, u, j) \end{aligned}$$

We assume the expectations to be well defined and finite. Note that

$\mathrm{E}[|g_t| \mid s_t = 0, u_t = 1] = 0$ and $p_{00}(1) = 1$; i.e. state 0 is a zero cost absorption state with $\mathcal{A}(0) = \{1\}$.

Let $h_t = (s_0, u_0, g_0, s_1, u_1, g_1, \ldots, s_{t-1}, u_{t-1}, g_{t-1}, s_t)$ denote the history of the process upto time t, where $t \in \mathbf{N}_0$, with $h_0 = (s_0)$. The history h_t follows the

recursion $h_t = (h_{t-1}, u_{t-1}, g_{t-1}, s_t)$ for $t \geq 1$. Let $\mathcal{H}_t$ denote the set of histories upto time t. $\mathcal{H}_0 = \mathcal{S}$, $\mathcal{H}_{t+1} = \mathcal{H}_t \mathcal{A}\mathbf{R}\mathcal{S}$ for $t \geq 0$. The sample space $\Omega = \mathcal{H}^\infty = (\mathcal{S}\mathcal{A}\mathbf{R})^\infty$ is the set of all infinite sequences of the form $(s_0, u_0, g_0, s_1, u_1, g_1, \ldots, s_t, u_t, g_t, \ldots)$, where $s_t \in \mathcal{S}, u_t \in \mathcal{A}, g_t \in \mathbf{R}$. This space is endowed with the product topology. Here $\mathcal{S}$ and $\mathcal{A}$ are endowed with the discrete topology and the real line $\mathbf{R}$ is endowed with the Borel topology.

An admissible or feasible policy ν for the SSP is a sequence of stochastic control kernels ν_t on $\mathcal{A}$, (i.e. $\nu = (\nu_0, \nu_1, \nu_2, \ldots)$) given the past history h_t, with the restriction that $\nu_t(\mathcal{A}(s_t) \mid h_t) = 1$; i.e. the probability measure should be concentrated on the set of feasible actions. $\mathcal{M}$ denotes the set of all feasible policies. Let $\mathcal{P}^\nu(\cdot \mid i)$ denote the probability measure induced on Ω under policy ν, starting from state $s_0 = i$, where $i \in \mathcal{S}$. $\mathrm{E}^\nu(\cdot \mid i)$ denotes the corresponding expectation, under the probability measure induced by policy ν, starting from state $s_0 = i$, where $i \in \mathcal{S}$. For the definition of Markov Randomized policy, Markov Deterministic policy and Stationary policy see Chapter 1.

For SSP problems, in the case of Markov policies we implicitly assume without loss of generality that in the termination state '0', the action taken is always the unique action '1'.

See Chapter 4 for notations on stationary randomized policies. The set of stationary randomized policies (or stochastic control kernels to be precise) is denoted by Λ. $\delta \in \Lambda$ may be used to represent the stochastic kernel or the stationary randomized policy; it will be clear from the context what we mean. $[\delta(i)]_a$ for $i \in \{1, 2, \ldots, n\}$ and $a \in \mathcal{A}(i)$ denotes the probability of taking action a from state

i under the control kernel δ.

The set of stationary deterministic policies (or control functions to be precise) is denoted by Υ. For $\mu \in \Upsilon$, $\mu(i) \in \mathcal{A}(i)$ denotes the action taken from state i for $i \in \{1, 2, \ldots, n\}$. The cardinality of Υ is

$$|\Upsilon| = |\mathcal{A}(1)| \times |\mathcal{A}(2)| \times \cdots \times |\mathcal{A}(n)|$$

Note that $\mu \in \Upsilon$ may be used to represent the control function or the stationary deterministic policy; what we mean will be clear from the context.

The cost to go function for the SSP problem for policy $\nu \in \mathcal{M}$, starting from state $i \in \{1, 2, \ldots, n\}$ is defined as

$$\tilde{J}^{\nu}(i) = \limsup_{k \to \infty} \mathrm{E}^{\nu} \left[\sum_{t=0}^{k} g_t \mid s_0 = i \right]$$

$\tilde{J}^{\nu} \in \mathbf{R}^n$ denotes the cost to go vector. See Chapter 1 for more details and the conditions under which the limit exist (instead of lim sup) in the above definition. Note that we use $\Pr^{\nu}$ interchangeably for $\mathcal{P}^{\nu}$.

C.2.1 Non-Termination Probability Of SSP MDPs

We are interested in finding the k stage non-termination (non-absorption) probability for the SSP problem.

Now for $k \in \mathbf{N}_0$ and a feasible policy $\nu = (\nu_0, \nu_1, \nu_2, \ldots) \in \mathcal{M}$

$$\Pr^{\nu} [s_k \neq 0 \mid s_0 = i] = \mathrm{E}^{\nu} \left[\boldsymbol{\mathcal{I}}_{[s_k \neq 0]} \mid s_0 = i \right]$$

for $i \in \mathcal{S}$. Here $\Pr^{\nu}$ denotes the probability distribution induced under policy ν, and likewise E^{ν} represents the expectation under policy ν. $\boldsymbol{\mathcal{I}}$ denotes the indicator

function. Notice that in determining $\Pr^{\nu}[s_k \neq 0 \mid s_0 = i]$, $i \in \{1, 2, \ldots, n\}$ only the decisions taken in the first k stages are relevant, i.e. only the stochastic control kernels $\nu_0, \nu_1, \ldots, \nu_{k-1}$ are relevant. Observe that $\Pr^{\nu}[s_0 \neq 0 \mid s_0 = i] = 1$ for $i \in \{1, 2, \ldots, n\}$.We are interested in finding

$$\sup_{\nu \in \mathcal{M}} \Pr^{\nu}[s_k \neq 0 \mid s_0 = i]$$

for $i \in \{1, 2, \ldots, n\}$. By dynamic programming argument (see later subsection) we can see that there exists a k-stage Markov deterministic policy $(\mu_0^k, \mu_1^k, \ldots, \mu_{k-1}^k)$, where $\mu_t^k, t \in \{0, 1, 2, \ldots, k-1\}$ is the control function used at time t, that maximizes the above probability for all $i \in \{1, 2, \ldots, n\}$. Let $\nu^k = (\nu_0^k, \nu_1^k, \ldots)$ be a feasible policy such that ν_t^k "equals" μ_t^k for $t = 0, 1, \ldots, k-1$ and ν_t^k arbitrary for $t \geq k$. Then

$$\Pr^{\nu^k}[s_k \neq 0 \mid s_0 = i] = \sup_{\nu \in \mathcal{M}} \Pr^{\nu}[s_k \neq 0 \mid s_0 = i]$$

for $i \in \{1, 2, \ldots, n\}$.

For $k \in \mathbf{N}_0$ and $\nu \in \mathcal{M}$ define

$$\rho_{\nu,k} \equiv \max_{i \in \{1,2,\ldots,n\}} \Pr^{\nu}[s_k \neq 0 \mid s_0 = i] \tag{C.3}$$

Since state 0 is an absorption state (i.e. it remains in state 0 once it is reached) we have

$$\Pr^{\nu}[s_{k+1} \neq 0 \mid s_0] \leq \Pr^{\nu}[s_k \neq 0 \mid s_0 = i]$$

for $k \geq 0$ and $i \in \{1, 2, \ldots, n\}$. Hence $\rho_{\nu,k}$ is a nonincreasing function for any fixed $\nu \in \mathcal{M}$, i.e. $\rho_{\nu,k} \downarrow_k$.

In this section, for stationary randomized policy (actually stochastic control kernel) $\delta \in \Lambda$, let P_δ denote the $n \times n$ substochastic matrix with

$$[P_\delta]_{ij} = \sum_{a \in \mathcal{A}(i)} [\delta(i)]_a \, p_{ij}(a)$$

for $i, j \in \{1, 2, \ldots, n\}$. In particular for stationary deterministic policy (actually control function) $\mu \in \Upsilon$

$$[P_\mu]_{ij} = p_{ij}(\mu(i))$$

for $i, j \in \{1, 2, \ldots, n\}$.

Let policy ν be a Markov randomized policy for the SSP problem, where ν 'equal' to $(\delta_0, \delta_1, \ldots)$. Here $\delta_t \in \Lambda$ for $t \in \mathbf{N}_0$. Then for $i, j \in \{1, 2, \ldots, n\}$ and $t > 0$

$$\Pr^\nu [s_t = j \mid s_0 = i] = \left[P_{\delta_0} P_{\delta_1} \cdots P_{\delta_{t-1}} \right]_{ij}$$

$$\Pr^\nu [s_t \neq 0 \mid s_0 = i] = e_i^T \left(P_{\delta_0} P_{\delta_1} \cdots P_{\delta_{t-1}} \right) \underline{1}$$

where

$$e_i = [0, 0, \ldots, 0, \underbrace{1}_{i^{\text{th}} \text{ position}}, 0, \ldots, 0]^T$$

is the i^{th} co-ordinate vector (column vector) in $\mathbf{R}^n$ with one in the i^{th} position and zero elsewhere. $\underline{1} \in \mathbf{R}^n$ is the column vector with all entries equal to one; i.e.

$$\underline{1} = [1, 1, \ldots, 1]^T$$

Let us define for $k \in \mathbf{N}_0$

$$\tilde{\rho}_k \equiv \max_{\mu \in \Upsilon} \rho_{\mu,k} \tag{C.4}$$

$$\hat{\rho}_k \equiv \sup_{\nu \in \mathcal{M}} \rho_{\nu,k} \tag{C.5}$$

Let $k > 0$. With slight abuse of notation, we use ν to denote k stage Markov policies in the following. See that

$$\begin{aligned}\hat{\rho}_k &= \sup_{\substack{\nu=(\delta_0,\delta_1,\ldots,\delta_{k-1}) \\ \delta_t \in \Lambda}} \max_{i \in \{1,2,\ldots,n\}} \Pr^{\nu}[s_k \neq 0 \mid s_0 = i] \\ &= \max_{\substack{\nu=(\mu_0,\mu_1,\ldots,\mu_{k-1}) \\ \mu_t \in \Upsilon}} \max_{i \in \{1,2,\ldots,n\}} \Pr^{\nu}[s_k \neq 0 \mid s_0 = i]\end{aligned}$$

Observe that $\tilde{\rho}_k \downarrow_k$, $\hat{\rho}_k \downarrow_k$ and $\hat{\rho}_k \geq \tilde{\rho}_k$ for $k \geq 0$ with $\hat{\rho}_1 = \tilde{\rho}_1$ and $\hat{\rho}_0 = \tilde{\rho}_0 = 1$. For $\delta \in \Lambda$ define

$$p_{ij}^{\delta} \equiv \sum_{a \in \mathcal{A}(i)} [\delta(i)]_a \, p_{ij}(a)$$

for $i \in \{1, 2, \ldots, n\}$ and $j \in \mathcal{S}$. Note that $p_{00}^{\delta} = 1$ and $p_{0j}^{\delta} = 0$ for $j \in \{1, 2, \ldots, n\}$. p_{ij}^{δ} denotes the one stage transition probability for policy δ. We also have for $\mu \in \Upsilon$, $p_{ij}^{\mu} = p_{ij}(\mu(i))$ for $i \in \{1, 2, \ldots, n\}$ and $j \in \mathcal{S}$. Also $p_{00}^{\mu} = 1$ and $p_{0j}^{\mu} = 0$ for $j \in \{1, 2, \ldots, n\}$.

For $k \in \mathbf{N}_0$ and $\nu \in \mathcal{M}$, let $\hat{J}_k^{\nu} \in \mathbf{R}^n$ be such that

$$\hat{J}_k^{\nu}(i) = \mathrm{E}^{\nu}\left[\boldsymbol{\mathcal{I}}_{[s_k \neq 0]} \mid s_0 = i\right]$$

for $i \in \{1, 2, \ldots, n\}$. We are interested in finding

$$\sup_{\nu \in \mathcal{M}} \hat{J}_k^{\nu}(i) \qquad i \in \{1, 2, \ldots, n\}$$

Consider an associated problem in which the "immediate cost" is identically zero for any action from any state, but the transition probabilities remain the same as in the original SSP problem. Let the immediate cost at time $t \in \mathbf{N}_0$ be denoted by g_t, i.e.

$$\mathrm{E}[|\mathsf{g}_t| \mid s_t = i, u_t = a] = 0 \qquad i \in \mathcal{S}, a \in \mathcal{A}(i) \text{ and } t \in \mathbf{N}_0$$

Hence

$$\begin{aligned} \mathsf{g}(i,a,j) &\equiv \mathrm{E}\left[\mathsf{g}_t \mid s_t = i, u_t = a, s_{t+1} = j\right] \\ &= 0 \end{aligned}$$

for $i, j \in \mathcal{S}$ and $a \in \mathcal{A}(i)$. Also for $i \in \mathcal{S}$, $a \in \mathcal{A}(i)$

$$\begin{aligned} \mathsf{g}(i,a) &\equiv \mathrm{E}\left[\mathsf{g}_t \mid s_t = i, u_t = a\right] \\ &= 0 \end{aligned}$$

For $\delta \in \Lambda$, the expected immmediate cost vector $\bar{\mathsf{g}}^{\delta} \in \mathbf{R}^n$ is given by

$$\begin{aligned} \bar{\mathsf{g}}^{\delta}(i) &= \sum_{a \in \mathcal{A}(i)} [\delta(i)]_a \, \mathsf{g}(i,a) \\ &= 0 \end{aligned}$$

for $i \in \{1, 2, \ldots, n\}$. Similarly for $\mu \in \Upsilon$, the expected "immediate cost" vector $\bar{\mathsf{g}}^{\mu} \in \mathbf{R}^n$ is given by

$$\begin{aligned} \bar{\mathsf{g}}^{\mu}(i) &= \mathsf{g}(i, \mu(i)) \\ &= 0 \end{aligned}$$

for $i \in \{1, 2, \ldots, n\}$. That is $\bar{\mathsf{g}}^{\delta}$ and $\bar{\mathsf{g}}^{\mu}$ are zero vectors. Evaluating $\hat{J}_k^{\nu}$ corresponds to a k stage problem with identically zero "immediate costs" and a terminal cost of one if $s_k \neq 0$ and zero if $s_k = 0$. Note that for policy ν the history used is the same as the history of the original problem. For instance if past immediate costs are included in the history for taking decisions, the immediate cost of the original SSP problem is the one which is used.

Now for the associated problem, define operators $\hat{T}_\delta, \hat{T}_\mu$ and $\hat{T}$ from $\mathbf{R}^n$ to $\mathbf{R}^n$, for $\delta \in \Lambda, \mu \in \Upsilon$ as follows. For $J \in \mathbf{R}^n$

$$\begin{aligned} \hat{T}_\delta J &= \bar{\mathsf{g}}^\delta + P_\delta J \\ &= P_\delta J \\ \hat{T}_\mu J &= \bar{\mathsf{g}}^\mu + P_\mu J \\ &= P_\mu J \end{aligned}$$

Let

$$\begin{aligned} \left(\hat{T}J\right)(i) &= \max_{a \in \mathcal{A}(i)} \left(\mathsf{g}(i,a) + \sum_{j=1}^{n} p_{ij}(a)\, J(j) \right) \qquad \text{for } i \in \{1,2,\ldots,n\} \\ &= \max_{a \in \mathcal{A}(i)} \left(\sum_{j=1}^{n} p_{ij}(a)\, J(j) \right) \end{aligned}$$

i.e.

$$\hat{T}J = \max_{\mu \in \Upsilon} \hat{T}_\mu J$$

where the maximization is taken componentwise over each index $i \in \{1,2,\ldots,n\}$. With $\underline{1} \in \mathbf{R}^n$ being the vector with all components equal to one, let

$$\hat{J}_k^* = \hat{T}^k \underline{1}$$

for $k \in \mathbf{N}_0$. Here $\hat{T}^k = \hat{T}\hat{T}^{k-1}$ is the composition of $\hat{T}$ with itself k times. $\hat{T}^0$ is the identity operator. Hence $\hat{J}_0^* = \underline{1}$. Since state '0' is the zero cost absorbing state we have

$$\mathrm{E}^\nu \left[\boldsymbol{\mathcal{I}}_{[s_k \neq 0]} \mid s_l = 0 \right] = 0 \qquad \text{for } l = 0, 1, \ldots k;\ \nu \in \mathcal{M}$$

From the dynamic programming argument we get

$$\hat{J}_k^*(i) = \sup_{\nu \in \mathcal{M}} \mathrm{Pr}^\nu \left[s_k \neq 0 \mid s_0 = i \right]$$

for $i \in \{1, 2, \ldots, n\}$ and $k \in \mathbf{N}_0$. Let

$$\hat{\mu}_k = \arg\max_{\mu \in \Upsilon} \hat{T}_\mu \hat{J}_k^*$$

i.e. $\hat{T}_{\hat{\mu}_k} \hat{J}_k^* = \hat{T} \hat{J}_k^*$.

Fix $k \in \{1, 2, \ldots\}$. Let $\hat{\nu}^k = (\hat{\nu}_0^k, \hat{\nu}_1^k, \hat{\nu}_2^k, \ldots)$ be an admissible policy such that $\hat{\nu}_l^k$ 'equal' to $\hat{\mu}_{k-1-l}$, for $l = 0, 1, \ldots, k-1$. $\hat{\nu}_l^k$ is arbitrary for $l \geq k$. (Note the slight abuse of notation; since $\hat{\nu}_l^k$ is a stochastic kernel on $\mathcal{A}$ given history h_l with the restriction that $\hat{\nu}_l^k(\mathcal{A}(s_l) \mid h_l) = 1$, while $\hat{\mu}_{k-1-l}$ is a control function.)

Thus

$$\Pr^{\hat{\nu}^k} [s_k \neq 0 \mid s_0 = i] = \hat{J}_k^*(i)$$

for $i \in \{1, 2, \ldots, n\}$. Note that

$$\hat{\rho}_k \quad = \quad \max_{i \in \{1,2,\ldots,n\}} \hat{J}_k^*(i)$$

for $k \in \mathbf{N}_0$.

The following example shows that $\hat{\rho}_k$ can be strictly greater than $\tilde{\rho}_k$ for $k > 1$.

Example C.3 *Consider a homogeneous* SSP *problem with states* $\mathcal{S} = \{0, 1, 2\}$*, with '*$0$*' being the termination state. Let the control constraints be* $\mathcal{A}(0) = \{1\}$, $\mathcal{A}(1) = \{1, 2\}$, $\mathcal{A}(2) = \{1\}$. *The immediate cost of the problem is irrelevant since we are interested in finding* $\hat{\rho}_k$ *and* $\tilde{\rho}_k$ *for* $k \in \mathbf{N}_0$.

Let the transition probabilities be $p_{11}(1) = \frac{1}{2}$, $p_{10}(1) = \frac{1}{2}$; $p_{12}(2) = \frac{2}{3}$, $p_{10}(2) = \frac{1}{3}$; $p_{20}(1) = 1$; $p_{00}(1) = 1$. *Let*

$$\bar{\mu} = \begin{pmatrix} 1 \\ \\ 1 \end{pmatrix} \qquad \textit{and} \qquad \tilde{\mu} = \begin{pmatrix} 2 \\ \\ 1 \end{pmatrix}$$

be the two possible Markov Deterministic control functions. Adhering to the notations in this section, let $\underline{1} = (1\ 1)'$ *be the two dimensional column vector with all the elements equal to one. Then*

$$\hat{J}_0^{\bar{\mu}} = \hat{J}_0^{\tilde{\mu}} = \begin{pmatrix} 1 \\ \\ 1 \end{pmatrix}$$

Also we have

$$\begin{aligned}
\hat{J}_1^{\bar{\mu}} &= \begin{pmatrix} \frac{1}{2} \\ \\ 0 \end{pmatrix} \\
\hat{J}_k^{\bar{\mu}} &= \begin{pmatrix} \frac{1}{2^k} \\ \\ 0 \end{pmatrix} \qquad \textit{for } k \geq 1 \\
\hat{J}_1^{\tilde{\mu}} &= \begin{pmatrix} \frac{2}{3} \\ \\ 0 \end{pmatrix} \\
\hat{J}_k^{\tilde{\mu}} &= \begin{pmatrix} 0 \\ \\ 0 \end{pmatrix} \qquad \textit{for } k > 1
\end{aligned}$$

Here

$$\hat{J}_k^{\mu}(i) = \mathrm{E}^{\mu}\left[\boldsymbol{\mathcal{I}}_{[s_k \neq 0]} \mid s_0 = i\right] = \left(\hat{T}_{\mu}^k \underline{1}\right)(i)$$

for $i \in \{1, 2\},\ \mu \in \{\bar{\mu}, \tilde{\mu}\},\ k \in \mathbf{N}_0$. *Hence we have*

$$\begin{aligned}
\tilde{\rho}_1 &= \frac{2}{3} \\
\tilde{\rho}_k &= \frac{1}{2^k} \qquad k \geq 2
\end{aligned}$$

Now $\hat{J}_k^* = \hat{T}^k \underline{1}$ *for* $k \in \mathbf{N}_0$. *Also*

$$\hat{J}_k^*(i) = \sup_{\nu \in \mathcal{M}} \mathrm{E}^{\nu}\left[\boldsymbol{\mathcal{I}}_{[s_k \neq 0]} \mid s_0 = i\right]$$

for $i \in \{1,2\}$ *and* $k \in \mathbf{N}_0$. *We have*

$$\hat{J}_0^* = \begin{pmatrix} 1 \\ \\ 1 \end{pmatrix}$$

$$\hat{J}_1^* = \begin{pmatrix} \frac{2}{3} \\ \\ 0 \end{pmatrix}$$

$$\hat{J}_k^* = \begin{pmatrix} \frac{2}{3}\frac{1}{2^{k-1}} \\ \\ 0 \end{pmatrix} \qquad \textit{for } k \geq 2$$

Note that $\hat{\mu}_0 = \tilde{\mu}$ *and* $\hat{\mu}_k = \bar{\mu},\ k \geq 1$ *where*

$$\hat{T}_{\hat{\mu}_k} \hat{J}_k^* = \hat{T} \hat{J}_k^*$$

For $k = 1$, *the policy which uses* $\tilde{\mu}$ *at stage* 0, *maximizes the one stage non-termination probability. For* $k > 1$, *the policy which uses* $\bar{\mu}$ *for the first* $k-1$ *stages (from stages* 0 *to* $k-2$*) and* $\tilde{\mu}$ *at stage* $k-1$ *is the one which maximizes the* k *stage non-termination probability.*

Note that

$$\hat{\rho}_k = \frac{2}{3}\frac{1}{2^{k-1}} \qquad \textit{for } k \geq 1$$

$$\frac{\hat{\rho}_k}{\tilde{\rho}_k} = \frac{\frac{2}{3}\frac{1}{2^{k-1}}}{\frac{1}{2^k}} = \frac{4}{3} > 1 \qquad \textit{for } k > 1$$

□

C.2.2 Absorption Or Termination Probability Of SSP MDPs

In this subsection we consider an associated problem to the original SSP problem which can be considered to be the complementary part of the results in the

previous subsection.

We are interested in finding the k-stage termination or absorption probability of the SSP problem.

Now for $k \in \mathbf{N}_0$ and any feasible policy $\nu = (\nu_0, \nu_1, \nu_2, \ldots) \in \mathcal{M}$

$$\Pr^{\nu}[s_k = 0 \mid s_0 = i] = \mathrm{E}^{\nu}\left[\boldsymbol{\mathcal{I}}_{[s_k=0]} \mid s_0 = i\right]$$

Here $\Pr^{\nu}$ denotes the induced probability under policy ν and E^{ν} is the expectation under policy ν (given starting state $s_0 = i$). $\boldsymbol{\mathcal{I}}$ denotes the indicator function. Note that

$$\Pr^{\nu}[s_0 = 0 \mid s_0 = i] = 0 \qquad i \in \{1, 2, \ldots, n\}$$

and $(\Pr^{\nu}[s_k = 0 \mid s_0 = i]) \uparrow_k$ for fixed state $i \in \mathcal{S}$ and $\nu \in \mathcal{M}$, since state '0' is a self absorption state. Notice that in determining $\Pr^{\nu}[s_k = 0 \mid s_0 = i]\,,\ i \in \{1, 2, \ldots, n\}$, only the decisions taken in the first k stages are relevant, i.e. only the stochastic control kernels $\nu_0, \nu_1, \ldots, \nu_{k-1}$ are relevant.

For $k \in \mathbf{N}_0$ and $\nu \in \mathcal{M}$ let $\check{J}_k^{\nu} \in \mathbf{R}^n$ be such that

$$\begin{aligned} \check{J}_k^{\nu}(i) &= \mathrm{E}^{\nu}\left[\boldsymbol{\mathcal{I}}_{[s_k=0]} \mid s_0 = i\right] \\ &= \Pr^{\nu}[s_k = 0 \mid s_0 = i] \end{aligned}$$

for $i \in \{1, 2, \ldots, n\}$.

We are interested in finding

$$\inf_{\nu \in \mathcal{M}} \Pr^{\nu}[s_k = 0 \mid s_0 = i] = \inf_{\nu \in \mathcal{M}} \check{J}_k^{\nu}(i)$$

for $i \in \{1, 2, \ldots, n\}$. Let $P_\delta,\ \delta \in \Lambda$ and $P_\mu,\ \mu \in \Upsilon$ be defined as in the previous

subsection. Note that

$$\check{J}_k^\nu(i) = 1 - \hat{J}_k^\nu(i)$$

where

$$\hat{J}_k^\nu(i) \equiv \mathrm{E}^\nu \left[\boldsymbol{\mathcal{I}}_{[s_k \neq 0]} \mid s_0 = i \right]$$

for $i \in \{1, 2, \ldots, n\}$ as defined in the previous subsection.

Consider the associated problem in which the "immediate costs" are as follows. g_t denotes the immediate cost at time $t \in \mathbf{N}_0$.

$$\mathrm{E}\left[|\mathsf{g}_t| \mid s_t = i, u_t = a, s_{t+1} = j \right] = 0$$

for $i, j \in \{1, 2, \ldots, n\}$ and $a \in \mathcal{A}(i)$.

$$\mathrm{E}\left[|\mathsf{g}_t - 1| \mid s_t = i, u_t = a, s_{t+1} = 0 \right] = 0$$

for $i \in \{1, 2, \ldots, n\}$ and $a \in \mathcal{A}(i)$ and

$$\mathrm{E}\left[|\mathsf{g}_t| \mid s_t = 0, u_t = 1, s_{t+1} = j \right] = 0$$

for $j \in \mathcal{S}$.

Assume that the transition probabilities remain the same as in the original SSP problem. That is we assume that a unit cost is incurred when state $s_t \in \{1, 2, \ldots, n\}$ and state $s_{t+1} = 0$, and a zero cost otherwise. Hence

$$\mathsf{g}(i, u, j) \equiv \mathrm{E}\left[\mathsf{g}_t \mid s_t = i, u_t = u, s_{t+1} = j \right] = 0$$

if $i = 0$ or $j \neq 0$. Here $u \in \mathcal{A}(i)$. Also

$$\mathsf{g}(i, u, 0) \equiv \mathrm{E}\left[\mathsf{g}_t \mid s_t = i, u_t = u, s_{t+1} = 0 \right] = 1$$

if $i \in \{1, 2, \ldots, n\},\ u \in \mathcal{A}(i)$. Note that

$$\begin{aligned} \mathrm{g}(i,u) &\equiv \mathrm{E}\left[\mathrm{g}_t \mid s_t = i, u_t = u\right] \\ &= \sum_{j=0}^{n} p_{ij}(u)\, \mathrm{g}(i,u,j) \\ &= p_{i0}(u) \qquad \text{for } i \in \{1, 2, \ldots, n\},\ u \in \mathcal{A}(i) \end{aligned}$$

Also

$$\mathrm{g}(0,1) \equiv \mathrm{E}\left[\mathrm{g}_t \mid s_t = 0, u_t = 1\right] = 0$$

If the "immediate cost" vector for $\delta \in \Lambda$ is denoted by $\bar{\mathrm{g}}^{\delta} \in \mathbf{R}^n$, then

$$\bar{\mathrm{g}}^{\delta}(i) = \sum_{a \in \mathcal{A}(i)} [\delta(i)]_a\, \mathrm{g}(i,a) = \sum_{a \in \mathcal{A}(i)} [\delta(i)]_a\, p_{i0}(a)$$

for $i \in \{1, 2, \ldots, n\}$ Similarly for $\mu \in \Upsilon$, the immediate cost vector $\bar{\mathrm{g}}^{\mu} \in \mathbf{R}^n$ is such that

$$\bar{\mathrm{g}}^{\mu}(i) = \mathrm{g}(i, \mu(i)) = p_{i0}(\mu(i))$$

for $i \in \{1, 2, \ldots, n\}$.

For any feasible policy $\nu \in \mathcal{M}$ and $t \in \mathbf{N}_0$

$$\mathrm{E}^{\nu}\left[\mathrm{g}_t \mid s_0 = i\right] = \mathrm{Pr}^{\nu}\left[s_t \neq 0, s_{t+1} = 0 \mid s_0 = i\right]$$

and for integer $N > 0$

$$\mathrm{E}^{\nu}\left[\sum_{t=0}^{N-1} \mathrm{g}_t \mid s_0 = i\right] = \mathrm{Pr}^{\nu}\left[s_N = 0 \mid s_0 = i\right] = \check{J}_N^{\nu}(i)$$

for $i \in \{1, 2, \ldots, n\}$, since the termination state '0' is a zero cost absorption state.

Note that for policy $\nu \in \mathcal{M}$, the history used is the same as the history of the original SSP problem. For instance if past immediate costs are included in the

history for taking decisions, the immediate cost g_t of the original problem is the one which is used.

For Markov Randomized Policy $\nu = (\delta_0, \delta_1, \ldots)$ where $\delta_t \in \Lambda,\ t \in \mathbf{N}_0$, we have

$$\begin{aligned} \check{J}^{\nu}_{N+1} &= \bar{\mathbf{g}}^{\delta_0} + P_{\delta_0}\bar{\mathbf{g}}^{\delta_1} + P_{\delta_0}P_{\delta_1}\bar{\mathbf{g}}^{\delta_2} + \\ &\quad \ldots + P_{\delta_0}P_{\delta_1}\cdots P_{\delta_{N-1}}\bar{\mathbf{g}}^{\delta_N} \end{aligned}$$

for $N \in \mathbf{N}_0$.

Note that

$$P_{\delta_0}\underline{1} = \underline{1} - \bar{\mathbf{g}}^{\delta_0}$$

and

$$\begin{aligned} P_{\delta_0}P_{\delta_1}\cdots P_{\delta_N}\underline{1} &= \underline{1} - \Big[\bar{\mathbf{g}}^{\delta_0} + P_{\delta_0}\bar{\mathbf{g}}^{\delta_1} + P_{\delta_0}P_{\delta_1}\bar{\mathbf{g}}^{\delta_2} + \\ &\quad \ldots + P_{\delta_0}P_{\delta_1}\cdots P_{\delta_{N-1}}\bar{\mathbf{g}}^{\delta_N}\Big] \end{aligned} \tag{C.6}$$

This follows easily from the fact that

$$\hat{J}^{\nu}_k = \underline{1} - \check{J}^{\nu}_k \qquad \text{for } k \in \mathbf{N}_0$$

or by induction as follows.

$$P_{\delta_0}\underline{1} = \underline{1} - \bar{\mathbf{g}}^{\delta_0}$$

is straightforward. Suppose equation C.6 is true for $N \in \mathbf{N}_0$. Since $P_{\delta_k}\underline{1} = \underline{1} - \bar{\mathbf{g}}^{\delta_k}$ for all $k \in \mathbf{N}_0$, we have

$$\underline{1} - \Big[\Big(\bar{\mathbf{g}}^{\delta_0} + P_{\delta_0}\bar{\mathbf{g}}^{\delta_1} + P_{\delta_0}P_{\delta_1}\bar{\mathbf{g}}^{\delta_2} + \ldots + P_{\delta_0}P_{\delta_1}\cdots P_{\delta_{N-1}}\bar{\mathbf{g}}^{\delta_N}\Big) + P_{\delta_0}P_{\delta_1}\cdots P_{\delta_N}\bar{\mathbf{g}}^{\delta_{N+1}}\Big]$$

$$= P_{\delta_0}P_{\delta_1}\cdots P_{\delta_N}\underline{1} - P_{\delta_0}P_{\delta_1}\cdots P_{\delta_N}\bar{\mathbf{g}}^{\delta_{N+1}}$$

$$= P_{\delta_0} P_{\delta_1} \cdots P_{\delta_N} \left[\underline{1} - \bar{\mathsf{g}}^{\delta_{N+1}} \right]$$

$$= P_{\delta_0} P_{\delta_1} \cdots P_{\delta_N} P_{\delta_{N+1}} \underline{1}$$

Define the operators $\check{T}_\delta, \check{T}_\mu$ from $\mathbf{R}^n$ to $\mathbf{R}^n$ for $\delta \in \Lambda,\ \mu \in \Upsilon$ by

$$\check{T}_\delta J = \bar{\mathsf{g}}^\delta + P_\delta J \qquad \text{for } J \in \mathbf{R}^n$$

$$\check{T}_\mu J = \bar{\mathsf{g}}^\mu + P_\mu J \qquad \text{for } J \in \mathbf{R}^n$$

Let the operator $\check{T} : \mathbf{R}^n \to \mathbf{R}^n$ be defined by

$$\left(\check{T} J\right)(i) = \min_{a \in \mathcal{A}(i)} \left[\mathsf{g}(i,a) + \sum_{j=1}^{n} p_{ij}(a)\ J(j) \right]$$

for $i \in \{1, 2, \ldots, n\}$; i.e.

$$\check{T} J = \min_{\mu \in \Upsilon} \check{T}_\mu J$$

where the minimization is taken componentwise.

Let $\underline{0} \in \mathbf{R}^n$ be the zero vector (column vector) with all components equal to zero. Let

$$\check{J}_k^* = \check{T}^k \underline{0} \qquad \text{for } k \in \mathbf{N}_0$$

where $\check{T}^k$ is equal to $\check{T}\check{T}^{k-1}$, the composition of $\check{T}$ with itself k times. $\check{T}^0$ is the identity operator with $\check{J}_0^* = \underline{0}$.

From the Dynamic Programming argument

$$\check{J}_k^*(i) = \inf_{\nu \in \mathcal{M}} \Pr^\nu \left[s_k = 0 \mid s_0 = i \right]$$

for $i \in \{1, 2, \ldots, n\}$. Let

$$\check{\mu}_k = \arg \min_{\mu \in \Upsilon} \check{T}_\mu \check{J}_k^*$$

i.e.

$$\check{T}_{\check{\mu}_k}\check{J}_k^* = \check{T}\check{J}_k^*$$

Fix $k \in \{1, 2, \ldots\}$. Let $\check{\nu}^k = (\check{\nu}_0^k, \check{\nu}_1^k, \check{\nu}_2^k, \ldots)$ be an admissible policy such that $\check{\nu}_l^k$ 'equal to' $\check{\mu}_{k-1-l}$, for $l = 0, 1, 2, \ldots, k-1$ and $\check{\nu}_l^k$ arbitrary for $l \geq k$. Then

$$\Pr^{\check{\nu}^k}[s_k = 0 \mid s_0 = i] = \check{J}_k^*(i)$$

for $i \in \{1, 2, \ldots, n\}$. Note that since for any $\nu \in \mathcal{M}$ and $i \in \{1, 2, \ldots, n\}$, $(\Pr^{\nu}[s_k = 0 \mid s_0 = i]) \uparrow_k$ we have $\check{J}_k^*(i) \uparrow_k$ for any $i \in \{1, 2, \ldots, n\}$. Also

$$\hat{\rho}_k = 1 - \min_{i \in \{1,2,\ldots,n\}} \check{J}_k^*(i)$$

Let $k \in \mathbf{N}_0$ and $i \in \{1, 2, \ldots, n\}$. Since

$$\hat{J}_k^*(i) = \sup_{\nu \in \mathcal{M}} \Pr^{\nu}[s_k \neq 0 \mid s_0 = i]$$

and

$$\check{J}_k^*(i) = \inf_{\nu \in \mathcal{M}} \Pr^{\nu}[s_k = 0 \mid s_0 = i]$$

We have

$$\begin{aligned} \check{J}_k^*(i) &= \inf_{\nu \in \mathcal{M}} (1 - \Pr^{\nu}[s_k \neq 0 \mid s_0 = i]) \\ &= 1 - \sup_{\nu \in \mathcal{M}} \Pr^{\nu}[s_k \neq 0 \mid s_0 = i] \\ &= 1 - \hat{J}_k^* \end{aligned}$$

As an aside we have the following. Let $\hat{J} \in \mathbf{R}^n$ and $\check{J} = \underline{1} - \hat{J}$, where $\underline{1} \in \mathbf{R}^n$ is the column vector with all components equal to one. For $i \in \{1, 2, \ldots, n\}$ and

$a \in \mathcal{A}(i)$

$$\begin{aligned}
\sum_{j=1}^{n} p_{ij}(a)\ \hat{J}(j) &= \sum_{j=1}^{n} p_{ij}(a)\ \left(1 - \check{J}(i)\right) \\
&= 1 - p_{i0}(a) - \sum_{j=1}^{n} p_{ij}(a)\ \check{J}(j) \\
&= 1 - \left(p_{i0}(a) + \sum_{j=1}^{n} p_{ij}(a)\ \check{J}(j)\right) \qquad \text{(C.7)}
\end{aligned}$$

Hence for $\mu \in \Upsilon$

$$\hat{T}_{\mu}\hat{J} = \underline{1} - \check{T}_{\mu}\check{J}$$

Similarly for $\delta \in \Lambda$

$$\hat{T}_{\delta}\hat{J} = \underline{1} - \check{T}_{\delta}\check{J}$$

From equation C.7, we get

$$\begin{aligned}
\max_{a \in \mathcal{A}} \left(\sum_{j=1}^{n} p_{ij}(a)\ \hat{J}(j)\right) &= \max_{a \in \mathcal{A}(i)} \left[1 - \left(p_{i0}(a) + \sum_{j=1}^{n} p_{ij}(a)\ \check{J}(j)\right)\right] \\
&= 1 - \min_{a \in \mathcal{A}(i)} \left[p_{i0}(a) + \sum_{j=1}^{n} p_{ij}(a)\ \check{J}(j)\right]
\end{aligned}$$

for $i \in \{1, 2, \ldots, n\}$. Also $a^* \in \mathcal{A}(i)$ achieves the maximum on the left hand side if and only if a^* achieves the minimum in the minimization term on the right hand side of the previous equation. In vector notation

$$\begin{aligned}
\hat{T}\hat{J} &= \max_{\mu \in \Upsilon} \hat{T}_{\mu}\hat{J} \\
&= \underline{1} - \min_{\mu \in \Upsilon} \check{T}_{\mu}\check{J} \\
&= \underline{1} - \check{T}\check{J}
\end{aligned}$$

where the maximum and minimum are taken component wise. Also $\mu^* \in \Upsilon$ is such that $\hat{T}_{\mu^*}\hat{J} = \hat{T}\hat{J}$ if and only if $\check{T}_{\mu^*}\check{J} = \check{T}\check{J}$.

See the previous and the current subsections for the definition of $\hat{\mu}_k, \check{\mu}_k$ and $\hat{J}_k^*, \check{J}_k^*$ for $k \in \mathbf{N}_0$. Another way to prove that $\hat{J}_k^* = \underline{1} - \check{J}_k^*$ is given below.

Now $\hat{J}_0^* = \underline{1}$ and $\check{J}_0^* = \underline{0} = \underline{1} - \hat{J}_0^*$. Here $\underline{0} \in \mathbf{R}^n$ is the zero vector. Let $\hat{J}_k^* = \underline{1} - \check{J}_k^*$

Since

$$\hat{T}_{\hat{\mu}_k} \hat{J}_k^* = \hat{T} \hat{J}_k^*$$

we have

$$\check{T}_{\hat{\mu}_k} \left(\underline{1} - \hat{J}_k^* \right) = \check{T}_{\hat{\mu}_k} \check{J}_k^* = \check{T} \check{J}_k^*$$

Similarly since

$$\check{T}_{\check{\mu}_k} \check{J}_k^* = \check{T} \check{J}_k^*$$

we have

$$\hat{T}_{\check{\mu}_k} \left(\underline{1} - \check{J}_k^* \right) = \hat{T}_{\check{\mu}_k} \hat{J}_k^* = \hat{T} \hat{J}_k^*$$

Also

$$\begin{aligned} \hat{J}_{k+1}^* &= \hat{T} \hat{J}_k^* \\ &= \underline{1} - \check{T} \left(\underline{1} - \hat{J}_k^* \right) \\ &= \underline{1} - \check{T} \check{J}_k^* \\ &= \underline{1} - \check{J}_{k+1}^*. \end{aligned}$$

C.2.2.1 Notes On The Worst Case Non-Termination Probability Of SSP MDPs

Consider a homogeneous SSP problem. If all stationary deterministic policies are proper then $\hat{\rho}_n < 1$ (see [11, 12] and the future Section C.3 in this appendix). Hence for any SSP problem with all proper stationary deterministic policies, we have for any feasible policy $\nu \in \mathcal{M}$

$$\Pr^{\nu}[s_n \neq 0 \mid s_0 = i] \leq \hat{\rho}_n < 1$$

for all $i \in \{1, 2, \ldots, n\}$.

If $\nu = (\delta_0, \delta_1, \delta_2, \ldots)$ is any Markov Randomized policy where $\delta_t \in \Lambda$, is the stochastic control kernel used at stage t, then for l, k positive integers we have

$$\Pr^{\nu}[s_{l+k} \neq 0 \mid s_0 = i] \leq \hat{\rho}_l \, \hat{\rho}_k \qquad \text{for } i \in \{1, 2, \ldots, n\}$$

This follows easily from the Markov nature of the policy ν and the fact that state 0 is an absorption state.

If $\nu = (\nu_0, \nu_1, \nu_2, \ldots)$ is an arbitrary history dependent randomized feasible policy in $\mathcal{M}$ for the original SSP problem, then given any fixed starting state $s_0 = i$ where $i \in \mathcal{S}$, there exists (see Chapter 1 and [40, Chapter 5]) a Markov Randomized policy dependent on i and ν called, say $\bar{\nu} = (\delta_0, \delta_1, \delta_2, \ldots)$ where $\delta_t \in \Lambda$ for $t \in \mathbf{N}_0$. such that

$$\Pr^{\nu}[s_t = j \mid s_0 = i] = \Pr^{\bar{\nu}}[s_t = j \mid s_0 = i]$$

and

$$\Pr^{\nu}[s_t = j, u_t = a \mid s_0 = i] = \Pr^{\bar{\nu}}[s_t = j, u_t = a \mid s_0 = i]$$

for all $t \in \mathbf{N}_0,\ j \in \mathcal{S},\ a \in \mathcal{A}(j)$.

Hence for any feasible policy $\nu \in \mathcal{M}$ and positive integers l, k we have

$$\Pr^{\nu}\left[s_{l+k} \neq 0 \mid s_0 = i\right] \leq \hat{\rho}_l\, \hat{\rho}_k \qquad \text{for } i \in \{1, 2, \ldots, n\}$$

Thus

$$\max_{i \in \{1,2,\ldots,n\}} \sup_{\nu \in \mathcal{M}} \Pr^{\nu}\left[s_{l+k} \neq 0 \mid s_0 = i\right] \leq \hat{\rho}_l\, \hat{\rho}_k$$

i.e.

$$\hat{\rho}_{l+k} \leq \hat{\rho}_l\, \hat{\rho}_k$$

Hence for any feasible policy $\nu \in \mathcal{M}$

$$\Pr^{\nu}\left[s_{2n} \neq 0 \mid s_0 = i\right] \leq \hat{\rho}_n^2 \qquad \text{for } i \in \{1, 2, \ldots, n\}$$

and

$$\Pr^{\nu}\left[s_t \neq 0 \mid s_0 = i\right] \leq \hat{\rho}_n^{\lfloor \frac{t}{n} \rfloor} \qquad \text{for } i \in \{1, 2, \ldots, n\}$$

where $\lfloor \cdot \rfloor$ denotes the floor function.

Thus if all stationary deterministic policies are proper for the SSP problem, then

$$\sum_{t=0}^{\infty} \Pr^{\nu}\left[s_t \neq 0 \mid s_0 = i\right] < \infty \qquad \text{for } i \in \{1, 2, \ldots, n\}$$

As will be shown later, the above quantity is the expected number of steps for reaching the terminal state 0, under policy ν starting at state i at stage 0. Also note that

$$\sum_{t=0}^{\infty} \Pr^{\nu}\left[s_t \neq 0 \mid s_0 = 0\right] = 0$$

since state 0 is an absorbing state. This is the expected number of steps to reach the terminal state 0, starting at the terminal state 0.

C.2.3 Number Of Stages To Reach Terminal State

Suppose $\nu \in \mathcal{M}$ be a feasible or admissible policy for the original SSP problem. In this subsection we consider another associated SSP problem in which the transition probabilities are the same as in the original problem, but "immediate cost" g_t are as follows.

$$\mathrm{E}\left[|\mathsf{g}_t - 1| \mid s_t = i, u_t = a\right] = 0 \qquad \text{for } i \in \{1, 2, \ldots, n\},\ a \in \mathcal{A}(i)$$

and

$$\mathrm{E}\left[|\mathsf{g}_t| \mid s_t = 0, u_t = 1\right] = 0$$

Note that $\mathcal{A}(0) = \{1\}$. We have that the immediate cost of taking any action from any state in $\{1, 2, \ldots, n\}$ is 1, and the immediate cost of taking the action in the terminal state '0' is zero. Hence

$$\mathsf{g}(i, a) \equiv \mathrm{E}\left[\mathsf{g}_t \mid s_t = i, u_t = a\right] = 1 \qquad \text{for } i \in \{1, 2, \ldots, n\},\ a \in \mathcal{A}(i)$$

$$\mathsf{g}(0, 1) \equiv \mathrm{E}\left[\mathsf{g}_t \mid s_t = 0, u_t = 1\right] = 0$$

Note that $\mathsf{g}_t = \mathsf{g}(s_t, u_t)$. If N is the number of steps or stages required to reach the terminal state '0', then

$$N = \sum_{t=0}^{\infty} \mathsf{g}_t$$

since state '0' is a zero cost absorbing state. Note that

$$\mathrm{Pr}^{\nu}\left[s_t \neq 0 \mid s_0 = i\right] = \mathrm{E}^{\nu}\left[\mathsf{g}_t \mid s_0 = i\right]$$

for $i \in \mathcal{S}$. Here Pr^{ν} is the probability induced by policy ν and E^{ν} is the expectation under policy ν.

Note that for policy $\nu \in \mathcal{M}$, the history used is the same as the history of the original SSP problem. For instance if past immmediate costs are included in the history for taking decisions, the immmediate cost g_t of the original problem is the one which is used. Hence

$$\begin{aligned} \mathrm{E}^{\nu}\left[N \mid s_0 = i\right] &= \sum_{t=0}^{\infty} \mathrm{E}^{\nu}\left[\mathrm{g}_t \mid s_0 = i\right] \\ &= \sum_{t=0}^{\infty} \mathrm{Pr}^{\nu}\left[s_t \neq 0 \mid s_0 = i\right] \end{aligned}$$

for $i \in \mathcal{S}$. Note that for $t \in \mathbf{N}_0$,

$$\{N > t\} = \{s_t \neq 0\}$$

Hence

$$\mathrm{Pr}^{\nu}\left[s_t \neq 0 \mid s_0 = i\right] = \mathrm{Pr}^{\nu}\left[N \geq t+1 \mid s_0 = i\right]$$

for $i \in \mathcal{S}$.

Another way of looking at the expectation of N, or the average number of steps to reach terminal state is as follows. Since N is a non-negative integer valued random variable [21, pg. 42, Lemma 5.7; pg. 45, ex. 5.6]

$$\begin{aligned} \mathrm{E}^{\nu}\left[N \mid s_0 = i\right] &= \sum_{t=1}^{\infty} \mathrm{Pr}^{\nu}\left[N \geq t \mid s_0 = i\right] \\ &= \sum_{t=0}^{\infty} \mathrm{Pr}^{\nu}\left[s_t \neq 0 \mid s_0 = i\right] \end{aligned}$$

Note that this expected value is finite for all states $i \in \mathcal{S}$ if all stationary deterministic policies are proper.

In particular for a Markov Randomized policy $\nu = (\delta_0, \delta_1, \delta_2, \ldots)$ (where $\delta_t \in \Lambda$) and $i \in \{1, 2, \ldots, n\}$

$$\mathrm{E}^{\nu}\left[N \mid s_0 = i\right] = e_i^T \left[I + P_{\delta_0} + P_{\delta_0} P_{\delta_1} + P_{\delta_0} P_{\delta_1} P_{\delta_2} + \ldots\right] \underline{1}$$

which is finite if all stationary deterministic policies are proper. Here $e_i \in \mathbf{R}^n$ is the i^{th} co-ordinate vector whose i^{th} component is one and all other components are zero, $\underline{1} \in \mathbf{R}^n$ is the vector with all components equal to one and I is the $n \times n$ identity matrix (diagonal matrix with all diagonal entries equal to one).

For a 'proper' stationary deterministic policy δ and $i \in \{1, 2, \ldots, n\}$

$$\mathrm{E}^{\delta}\left[N \mid s_0 = i\right] = e_i^T \left(I - P_\delta\right)^{-1} \underline{1}$$

For a policy $\nu \in \mathcal{M}$ let $\bar{N}_\nu$ denote the $n \times 1$ vector with

$$\bar{N}_\nu(i) = \mathrm{E}^{\nu}\left[N \mid s_0 = i\right] \qquad \text{for } i \in \{1, 2, \ldots, n\}$$

Hence for 'proper' stationary randomized policy δ

$$\bar{N}_\delta = \left(I - P_\delta\right)^{-1} \underline{1}$$

Note that $\bar{N}_\nu(i) \geq 1$ for all $i \in \{1, 2, \ldots, n\}$.

C.3 Notes On The Non-Absorption Probability Of SSP MDPs

Note that $\rho_{\nu,k} \downarrow_k$ for any admissible policy $\nu \in \mathcal{M}$. Likewise $\tilde{\rho}_k \downarrow_k$ and $\hat{\rho}_k \downarrow_k$.

In the SSP problem the termination state '0' is a zero cost absorbing state. Hence for any stationary policy, whether randomized or deterministic, $\{0\}$ by itself is a recurrent class. A stationary policy $\delta \in \Lambda$ being 'proper' is equivalent to the statement that there are no recurrent states in $\{1, 2, \ldots, n\}$ under the policy δ, or equivalently the Markov Chain corresponding to δ has a single recurrent class (i.e. it is unichain) namely $\{0\}$. Let

$$p_{ij}^{\delta} \equiv \sum_{a \in \mathcal{A}(i)} \left[\delta(i)\right]_a \, p_{ij}(a)$$

be the one step transition probability under stationary policy δ, where $i, j \in \mathcal{S} = \{0, 1, 2, \ldots, n\}$. Note that for stationary deterministic policy $\mu \in \Upsilon$, $p_{ij}^{\mu} = p_{ij}(\mu(i))$ for $i \in \{1, 2, \ldots, n\}$, $j \in \mathcal{S}$ and $p_{00}^{\mu} = 1$. Also $p_{0j}^{\mu} = 0$ if $j \in \{1, 2, \ldots, n\}$.

Suppose $(s_0, s_1, \ldots, s_k)$ be a sequence of states with $k \geq 1$ and $s_l \in \mathcal{S}$ for $l \in \{0, 1, 2, \ldots, k\}$. This is called a path of length k with starting state s_0 and ending state s_k. If $s_0 = s_k$ we call this path a cycle.

We say that $(s_0, s_1, s_2, \ldots, s_k)$ is a path of positive probability under the stationary policy $\delta \in \Lambda$, if $p_{s_l s_{l+1}}^{\delta} > 0$ for $l = 0, 1, 2, \ldots, k-1$. Similarly if $(s_0, s_1, s_2, \ldots, s_k)$ is a cycle, we call it a cycle of positive probability under policy δ, if $p_{s_l s_{l+1}}^{\delta} > 0$ for $l = 0, 1, 2, \ldots, k-1$.

Now $(s_{k'}, s_{k'+1}, \ldots, s_{l'})$ with $0 \leq k' < l' \leq k$ is said to be a sub-path of $(s_0, s_1, s_2, \ldots, s_k)$ of length $l' - k'$. If $s_{l'} = s_{k'}$, then we say that $(s_{k'}, s_{k'+1}, \ldots, s_{l'})$ is a sub-cycle of the path $(s_0, s_1, s_2, \ldots, s_k)$. We say that the sub-path (sub-cycle) is a sub-path (sub-cycle) of positive probability under the stationary policy $\delta \in \Lambda$ if $p_{s_l s_{l+1}}^{\delta} > 0$ for $k' \leq l < l'$.

C.3.1 Properness Of Policies

Let $\delta \in \Lambda$. Note that if for some $k \in \mathbf{N}$ (where $\mathbf{N}$ is the set of positive integers) we have $\rho_{\delta,k} < 1$ then this implies that δ is proper. Also as will be shown below δ is proper if and only if $\rho_{\delta,k} < 1$ for all $k \geq n$.

δ proper implies that there exists $l \in \mathbf{N}$ such that $\rho_{\delta,l} < 1$. If $\rho_{\delta,l} < 1$ for some $l \leq n$ then $\rho_{\delta,n} < 1$ since $\rho_{\delta,k} \downarrow_k$.

If $\rho_{\delta,l} < 1$ for some $l \geq n$, then this implies that there is a path $(s_0, s_1, s_2, \ldots, s_l)$ of positive probability under δ of length l from each fixed starting state $i \in \mathcal{S} \setminus \{0\}$ to the terminal state 0; i.e. $s_0 = i, s_l = 0$ and $p^{\delta}_{s_k s_{k+1}} > 0$ for $0 \leq k < l$. Hence if we remove sub-cycles on this path , then there is a path of positive probability under δ , starting at state i and ending at state 0, of length $l' \leq n$ namely $(s'_0, s'_1, \ldots, s'_{l'})$ with $s'_0 = i,\ s'_{l'} = 0$ and $s'_k \neq s'_{k'}$ if $k' \neq k$ where $k, k' \in \{0, 1, \ldots, l'\}$. Also $p^{\delta}_{s'_k s'_{k+1}} > 0$ for $0 \leq k < l'$. Since 0 is a self absorbing state (i.e. $p^{\delta}_{00} = 1$) we can extend this path to a path$(s'_0, s'_1, \ldots, s'_{l'}, s'_{l'+1}, \ldots, s'_n)$ of positive probability under δ with $s'_k = 0$ for $l' < k \leq n$.

Since this is true for all $i \in \mathcal{S} \setminus \{0\}$, we have $\rho_{\delta,n} < 1$. Since $\rho_{\delta,k} \downarrow_k$ we have $\rho_{\delta,k} < 1$ for all $k \geq n$.

Now since we are dealing with a finite state, finite action SSP problem, there are only a finite number of stationary deterministic policies (i.e. $|\Upsilon|$ is finite). Now all stationary deterministic policies are proper is equivalent to the statement that the MDP is unichain [12, 40] with unique recurrent class $\{0\}$ for all stationary deterministic policies. Hence extending the above idea, if for some $k \in \mathbf{N}$ we have $\tilde{\rho}_k < 1$ then this implies that all stationary deterministic policies are proper. Also all stationary deterministic policies are proper if and only if $\tilde{\rho}_k < 1$ for all $k \geq n$.

Now if $\tilde{\rho}_k = 1$ for some $k \geq n$, then it is impossible that $\tilde{\rho}_l < 1$ for any $l \geq n$, since this would imply that all stationary deterministic policies are proper and hence $\tilde{\rho}_{l'} < 1$ for all $l' \geq n$ and hence $\tilde{\rho}_k < 1$; a contradiction. Hence for the SSP problem, either

$$\tilde{\rho}_k = 1 \qquad \text{for all } k \geq n, (\text{ and also } \tilde{\rho}_l = 1 \text{ for } 0 \leq l < n)$$

or

$$\tilde{\rho}_k < 1 \qquad \text{for all } k \geq n$$

Lemma C.4 *For any $k' \in \mathbf{N}_0$ we have $\tilde{\rho}_{k'} < 1$ if and only if $\hat{\rho}_{k'} < 1$, or equivalently $\tilde{\rho}_{k'} = 1$ if and only if $\hat{\rho}_{k'} = 1$*

□

Proof of Lemma C.4

For $k' = 0$ we have $\tilde{\rho}_0 = \hat{\rho}_0 = 1$. Hence we consider the case where $k' \geq 1$. Since $\hat{\rho}_k \geq \tilde{\rho}_k$ for all $k \in \mathbf{N}_0$ we have that $\hat{\rho}_{k'} < 1$ implies $\tilde{\rho}_{k'} < 1$. We will prove that $\tilde{\rho}_{k'} < 1$ implies that $\hat{\rho}_{k'} < 1$.

Note that starting at each initial state in $\mathcal{S} \setminus \{0\}$, the maximal non-termination probability for a k' stage problem is achieved by a Markov deterministic policy. Note that there are only finite number of k' stage Markov deterministic policies. Hence it is sufficient to prove that for any initial state $i \in \mathcal{S} \setminus \{0\}$ and any k' stage Markov Deterministic policy $\nu = (\mu_0, \mu_1, \ldots, \mu_{k'-1})$ (with slight abuse of notation we use ν to represent the k' stage policy) the k' stage termination probability is greater than zero. i.e.

$$\Pr^{\nu}[s_{k'} = 0 \mid s_0 = i] > 0 \qquad \text{for } i \in \mathcal{S} \setminus \{0\}$$

We will prove this as follows. Suppose not; i.e. there exists an $i \in \mathcal{S} \setminus \{0\}$ and k' stage Markov Deterministic policy $(\mu_0, \mu_1, \ldots, \mu_{k'-1})$ such that starting from state i at time zero (i.e. $s_0 = i$) the k' stage state $s_{k'}$ is not 0 with probability one.

Let

$$\mathcal{S}_0(i) = \{i\}$$

For $k = 0, 1, \ldots, k'$ let $\mathcal{S}_k(i)$ be defined as the set of all the states reachable in k steps or less, starting from state i, with positive probability under the Markov Deterministic policy $(\mu_0, \mu_1, \ldots, \mu_{k'-1})$. Note that

$$\mathcal{S}_k(i) \subseteq \mathcal{S}_{k+1}(i) \qquad \text{for } 0 \leq k < k'$$

Let $\mathcal{S}'_0(i) = \mathcal{S}_0(i)$ and for $k = 1, 2, \ldots, k'$ let

$$\mathcal{S}'_k(i) = \mathcal{S}_k(i) \setminus \mathcal{S}_{k-1}(i)$$

That is $\mathcal{S}'_k(i)$ is the set of states reachable with positive probability starting from state i, in k steps, but not in less than k steps under policy $(\mu_0, \mu_1, \ldots, \mu_{k'-1})$. Note that for $0 \leq k \leq k'$,

$$\mathcal{S}_k(i) = \cup_{l=0}^{k} \mathcal{S}'_l(i)$$

Note that $\mathcal{S}'_k(i) \cap \mathcal{S}'_l(i) = \emptyset$ for $k \neq l,\ k, l \in \{0, 1, 2, \ldots, k'\}$. $\mathcal{S}'_k(i)$ may be empty for some $k \in \{1, 2, \ldots, k'\}$. By assumption the terminal state $0 \notin \mathcal{S}_{k'}(i)$. Hence $0 \notin \mathcal{S}_k(i)$ for $0 \leq k \leq k'$. Let μ be a stationary deterministic policy (control function) such that

$$\mu(j) = \mu_l(j) \qquad \text{for } j \in \mathcal{S}'_l(i),\ 0 \leq l < k'$$

Let $\mu(j)$ be arbitrary for $j \notin \mathcal{S}_{k'-1}(i)$ with the restriction that $\mu(j) \in \mathcal{A}(j)$.

For $k \in \mathbf{N}_0$ let $\tilde{\mathcal{S}}_k(i)$ be defined as the set of all the states that can be reached in k steps or less with positive probability starting from state i under stationary deterministic policy μ. Note that $\tilde{\mathcal{S}}_0(i) = \mathcal{S}_0(i) = \{i\}$. We claim that

$$\tilde{\mathcal{S}}_k(i) \subseteq \mathcal{S}_k(i) \qquad \text{for } 0 \leq k \leq k'$$

Since $\mu(i) = \mu_0(i)$ we have $\mathcal{S}_1(i) = \tilde{\mathcal{S}}_1(i)$. For any $l \in \{0, 1, \ldots, k'-1\}$ and $j \in \mathcal{S}'_l(i)$, $p_{jj'}(\mu(j)) > 0$ implies $j' \in \mathcal{S}_{l+1}(i)$.

Suppose $\tilde{\mathcal{S}}_k(i) \subseteq \mathcal{S}_k(i)$ for some k with $0 \leq k < k'$. Now

$$\tilde{\mathcal{S}}_k(i) = \cup_{l=0}^{k} \left(\mathcal{S}'_l(i) \cap \tilde{\mathcal{S}}_k(i)\right)$$

Hence if $k \geq 1$, for $j \in \cup_{l=0}^{k-1} \left(\mathcal{S}'_l(i) \cap \tilde{\mathcal{S}}_k(i)\right)$, $p_{jj'}(\mu(j)) > 0$ implies $j' \in \mathcal{S}_k(i)$. Also $j \in \mathcal{S}'_k(i) \cap \tilde{\mathcal{S}}_k(i)$ and $p_{jj'}(\mu(j)) > 0$ implies $j' \in \mathcal{S}_{k+1}(i)$. Thus $\tilde{\mathcal{S}}_{k+1}(i) \subseteq \mathcal{S}_{k+1}(i)$.

Thus for $0 \leq k \leq k'$ we have

$$\tilde{\mathcal{S}}_k(i) \subseteq \mathcal{S}_k(i)$$

implying $0 \notin \tilde{\mathcal{S}}_{k'}(i)$. Thus $\rho_{\mu,k'} = 1$, a contradiction.

□

As an aside, see that if $\mathcal{S}_l(i) = \mathcal{S}_{l+1}(i)$ for some $l \in \{0, 1, \ldots, k'-1\}$, then $\tilde{\mathcal{S}}_k(i) \subseteq \mathcal{S}_l(i)$ for all $k \geq l$.

C.3.2 Acyclicity Of Policies

The following claim is self evident.

Claim C.1 *Assume that the states s_l, for $l = 0, 1, \ldots k$ in the path $(s_0, s_1, \ldots, s_k)$ be in $\mathcal{S} \setminus \{0\}$. If $k \geq n$ then at least one state in $\mathcal{S} \setminus \{0\}$ is repeated and hence there is a subcycle $(s_{k'}, s_{k'+1}, \ldots, s_{l'})$ of the above path, such that $s_{k'} = s_{l'}$ and $0 \leq k' < l' \leq k$. We can actually take $l' \leq n$.*

□

A stationary randomized policy $\delta \in \Lambda$ for the SSP problem is called acyclic if there are no cycles of positive probability (with states in $\mathcal{S} \setminus \{0\}$) under the policy δ.

Actually if there are no cycles of positive probability with states in $\mathcal{S} \setminus \{0\}$ under the policy δ of length less than or equal to n, then there are no cycles of positive probability with states in $\mathcal{S} \setminus \{0\}$ of length greater than or equal to n by the argument in Claim C.1.

Lemma C.5 *Suppose for some positive intiger k, $\rho_{\delta,k} = 0$. Then the Markov Chain under δ is acyclic.*

□

Proof of Lemma C.5

Suppose otherwise. That is δ is not acyclic. That means there is a cycle $(s_0, s_1, \ldots, s_{k'})$ of positive probability under δ with $s_l \in \mathcal{S} \setminus \{0\}$ for $l = 0, 1, \ldots, k'$ and $s_0 = s_{k'}$ (and $k' \leq n$) . Once we start within any of the states within this cycle at time 0, then we can remain in the states within this cycle with positive probability for any finite time. This implies $\rho_{\delta,k} > 0$, a contradiction.

□

Claim C.2 *Suppose $\delta \in \Lambda$ be acyclic. Then $\rho_{\delta,k} = 0$ for all $k \geq n$*

□

Proof of Claim C.2

Since $\rho_{\delta,l} \downarrow_l$ it is sufficient to prove that $\rho_{\delta,n} = 0$.

Suppose $\rho_{\delta,n} > 0$. Then there exists a starting state $s_0 = i \in \mathcal{S} \setminus \{0\}$ and a path $(s_0, s_1, \ldots, s_n)$ of positive probability under δ such that $s_l \in \mathcal{S} \setminus \{0\}$ for $l = 0, 1, \ldots, n$. This implies that by Claim C.1 there is a subcycle $(s_{k'}, s_{k'+1}, \ldots, s_{l'})$ of positive probability under δ with $0 \leq k' < l' \leq n$ and $s_{k'} = s_{l'}$. Hence δ is not acyclic, a contradiction. Thus $\rho_{\delta,k} = 0$ for all $k \geq n$ if δ is acyclic.

□

Since we are dealing with a finite state, finite action SSP problem, there are only a finite number of stationary deterministic policies; i.e. $|\Upsilon|$ is finite. Hence extending the above idea, if for some $k \in \mathbf{N}$, $\tilde{\rho}_k = 0$, then all stationary deterministic policies are acyclic. Also all stationary deterministic policies are acyclic implies that $\tilde{\rho}_k = 0$ for all $k \geq n$.

Now if $\tilde{\rho}_k > 0$ for some $k \geq n$, then it is not possible that $\tilde{\rho}_l = 0$ for any $l \geq n$ (and hence also for $0 \leq l < n$) since this implies that all stationary deterministic policies are acyclic and hence $\tilde{\rho}_{l'} = 0$ for all $l' \geq n$. Hence $\tilde{\rho}_k = 0$, a contradiction.

Hence for the SSP problem, either

$$\tilde{\rho}_k > 0 \qquad \text{for all } k \geq n, \text{ and also } \tilde{\rho}_k > 0 \text{ for } 0 \leq k < n$$

or

$$\tilde{\rho}_k = 0 \qquad \text{for all } k \geq n$$

Lemma C.6 *For any $k \in \mathbf{N}$ we have $\tilde{\rho}_k = 0$ if and only if $\hat{\rho}_k = 0$, or equivalently $\tilde{\rho}_k > 0$ if and only if $\hat{\rho}_k > 0$.* □

Proof of Lemma C.6

The fact that $\hat{\rho}_k = 0$ implies $\tilde{\rho}_k = 0$ follows easily. To prove the claim that $\tilde{\rho}_k = 0$ implies $\hat{\rho}_k = 0$, we need to observe that the worst case k stage non-absorption (or non-termination) probability for any starting state $i \in \mathcal{S} \setminus \{0\}$ is achieved by a Markov deterministic policy. Only the decisions in the first k stages are relevant and there are only a finite number of k stage Markov deterministic policies.

Suppose $\hat{\rho}_k > 0$. Then (with slight abuse of notation) there exists a k stage Markov deterministic policy $\nu = (\mu_0, \mu_1, \ldots, \mu_{k-1})$ (here $\mu_l \in \Upsilon$ for $l = 0, 1, \ldots, k-1$) such that for some starting state $s_0 = i \in \mathcal{S} \setminus \{0\}$ and ending state $s_k = j \in \mathcal{S} \setminus \{0\}$ we have a path $(s_0, s_1, \ldots, s_k)$ with$s_l \in \mathcal{S} \setminus \{0\}$ for $l = 0, 1, \ldots, k$ and $p_{s_l s_{l+1}}(\mu_l(s_l)) > 0$ for $l = 0, 1, \ldots, k-1$.

Case 1 : Suppose there are no sub cycles in the path $(s_0, s_1, \ldots, s_k)$, i.e. $s_l \neq s_{l'}$ for $l \neq l',\ l, l' \in \{0, 1, \ldots, k\}$. Note that this implies $k < n$. Suppose μ be a stationary deterministic policy such that $\mu(s_l) = \mu_l(s_l)$ for $l = 0, 1, \ldots, k-1$ and $\mu(s)$ arbitrary for the other states s, with the restriction that $\mu(s) \in \mathcal{A}(s)$. Hence we have a path $(s_0, s_1, \ldots, s_k)$ (with $s_0 = i,\ s_k = j$) of positive probability under stationary deterministic policy μ. Hence $\rho_{\mu,k} > 0$ implying $\tilde{\rho}_k > 0$, a contradiction.

Case 2: Suppose some state is repeated in the path $(s_0, s_1, \ldots, s_k)$ where $s_0 = i$ and $s_k = j$. Thus there is a sub cycle in the path. Hence there is a first sub cycle in the path (the sub cycle with the smallest terminal index) namely $(s_{l'}, s_{l'+1}, \ldots s_{k'})$ where $0 \leq l' < k' \leq k$. Here $s_{l'} = s_{k'}$ and $s_{l''} \neq s_{k''}$ for $l'' \neq k'',\ l'', k'' \in \{0, 1, \ldots, k'-1\}$. Note that $k' \leq n$.

For $s \in \{s_0, s_1, \ldots, s_{k-1}\}$ let

$$\hat{l}(s) = \arg \min_{l \in \{0,1,\ldots,k-1\}} s = s_l$$

For $s \in \{s_0, s_1, \ldots, s_{k-1}\}$ let us define

$$\mu(s) = \mu_{\hat{l}(s)}(s)$$

and let $\mu(s)$ be arbitrary with the restriction that $\mu(s) \in \mathcal{A}(s)$ for the other states.

Thus under stationary deterministic policy μ, $p_{s_l s_{l+1}}(\mu(s_l)) > 0$ for $l \in \{l', l'+1, \ldots, k'-1\}$. Thus $(s_{l'}, s_{l'+1}, \ldots, s_{k'})$ is a subcycle of postive probability under μ. Also if $l' > 0$ then $(s_0, s_1, \ldots, s_{l'})$ is a sub path of positive probability under μ. Hence $\rho_{\mu,l} > 0$ for all $l \geq 0$ implying that $\tilde{\rho}_l > 0$ for all $l \geq 0$, a contradiction.

□

Hence we may conclude that for any $k \in \mathbf{N}$

$$\hat{\rho}_k = 1 \iff \tilde{\rho}_k = 1$$

$$\hat{\rho}_k = 0 \iff \tilde{\rho}_k = 0$$

$$\hat{\rho}_k \in (0,1) \iff \tilde{\rho}_k \in (0,1)$$

Here $\iff$ stands for if and only if. From the above discussion we have the following.

For all stationary deterministic policies proper case, $\tilde{\rho}_l < 1$ or $\hat{\rho}_l < 1$ for some $l \in \mathbf{N}$ implying that $\tilde{\rho}_k < 1,\ \hat{\rho}_k < 1$ for all $k \geq n$ and all $k \geq l$.

If at least one of the stationary deterministic policies is not proper then $\hat{\rho}_l = 1$ or $\tilde{\rho}_l = 1$ for some $l \geq n$ implying $\tilde{\rho}_k = 1,\ \hat{\rho}_k = 1$ for all $k \in \mathbf{N}$.

If all stationary deterministic policies are acyclic, then $\hat{\rho}_l = 0$ or $\tilde{\rho}_l = 0$ for some $l \in \mathbf{N}$, implying $\hat{\rho}_k = \tilde{\rho}_k = 0$ for all $k \geq n$ and $k \geq l$.

If at least one stationary deterministic policy is not acyclic then $\hat{\rho}_l > 0$ or $\tilde{\rho}_l > 0$ for some $l \geq n$, implying $\hat{\rho}_k > 0,\ \tilde{\rho}_k > 0$ for all $k \in \mathbf{N}$.

Hence we have the following three cases.

Corresponding to at least one non proper stationary determinstc policy, we have

$$\hat{\rho}_k = 1,\ \tilde{\rho}_k = 1 \qquad \text{for all } k \in \mathbf{N}_0$$

Corresponding to all stationary deterministic policies acyclic, we have

$$\hat{\rho}_k = 0,\ \tilde{\rho}_k = 0 \qquad \text{for all } k \geq n$$

Corresponding to all stationary deterministic policies proper, but atleast one stationary deterministic policy is not acyclic, we have

$$\tilde{\rho}_k \in (0,1)\ \ \hat{\rho}_k \in (0,1) \qquad \text{for all } k \geq n$$

C.4 Contraction Properties Of SSP Dynamic Programming Operators

C.4.1 Preliminaries

Consider the following Dynamic Programming operators for SSP problem. For $\mu \in \Upsilon$ let $\tilde{T}_\mu : \mathbf{R}^n \to \mathbf{R}^n$ be such that, for $J \in \mathbf{R}^n$

$$\tilde{T}_\mu J = \bar{g}^\mu + P_\mu J$$

where $\bar{g}^\mu(i) = g(i, \mu(i)),\ i \in \{1, 2, \ldots, n\}$ and $[P_\mu]_{ij} = p_{ij}(\mu(i)),\ i, j \in \{1, 2, \ldots, n\}$. Similarly for $\delta \in \Lambda$, let $\tilde{T}_\delta : \mathbf{R}^n \to \mathbf{R}^n$ be such that, for $J \in \mathbf{R}^n$

$$\tilde{T}_\delta J = \bar{g}^\delta + P_\delta J$$

where

$$\bar{g}^{\delta}(i) = \sum_{a \in \mathcal{A}(i)} [\delta(i)]_a \, g(i,a)$$

and

$$[P_{\delta}]_{ij} = \sum_{a \in \mathcal{A}(i)} [\delta(i)]_a \, p_{ij}(a)$$

for $i, j \in \{1, 2, \ldots, n\}$. Here $g(i,a)$ is the expected immmediate cost of taking action a from state i, and $[\delta(i)]_a$ is the probability of taking action a from state i under stationary randomized policy (stochastic control kernel to be precise) δ. Note that $\bar{g}^{\mu}$ and $\bar{g}^{\delta}$ are expected immediate cost vectors, while P_{μ} and P_{δ} are $n \times n$ sub stochastic matrices.

Let the operator $\tilde{T} : \mathbf{R}^n \to \mathbf{R}^n$ be defined by

$$\tilde{T}J = \min_{\mu \in \Upsilon} \tilde{T}_{\mu} J$$

for $J \in \mathbf{R}^n$. Here the minimization is taken component wise, i.e.

$$\left(\tilde{T}J\right)(i) = \min_{a \in \mathcal{A}(i)} \left(g(i,a) + \sum_{j=1}^{n} p_{ij}(a) \; J(j) \right)$$

for $i \in \{1, 2, \ldots, n\}$.

It is easy to see that $\tilde{T}_{\mu}$ and $\tilde{T}_{\delta}$ are monotone operators for $\mu \in \Upsilon$ and $\delta \in \Lambda$. That is for $J, J' \in \mathbf{R}^n$, if $J' \geq J$ (i.e. $J'(i) \geq J(i)$ for $i \in \{1, 2, \ldots, n\}$), then

$$\begin{aligned} \tilde{T}_{\mu} J' &\geq \tilde{T}_{\mu} J \\ \tilde{T}_{\delta} J' &\geq \tilde{T}_{\delta} J \end{aligned}$$

where the inequality is component wise.

We will show that $\tilde{T}$ is a monotone operator too. For $J, J' \in \mathbf{R}^n$ let $J \leq J'$. Let $\mu \in \Upsilon$ be such that $\tilde{T}_\mu J' = \tilde{T} J'$. Hence

$$\tilde{T} J \leq \tilde{T}_\mu J \leq \tilde{T}_\mu J' = \tilde{T} J'$$

i.e. $\tilde{T} J \leq \tilde{T} J'$.

Note that for any $\delta \in \Lambda$ and $J, J' \in \mathbf{R}^n$

$$\tilde{T}_\delta (J + J') = \tilde{T}_\delta J + P_\delta J'$$

Let $J \in \mathbf{R}^n$ and $\epsilon \geq 0,\ \epsilon \in \mathbf{R}$. For $k \geq 1$, k integer and $\delta_0, \delta_1, \delta_2, \ldots, \delta_{k-1} \in \Lambda$, we have

$$\begin{aligned} \tilde{T}_{\delta_0} \tilde{T}_{\delta_1} \cdots \tilde{T}_{\delta_{k-1}} (J + \epsilon\, \underline{1}) &= \tilde{T}_{\delta_0} \tilde{T}_{\delta_1} \cdots \tilde{T}_{\delta_{k-1}} J + \epsilon \left(\prod_{l=0}^{k-1} P_{\delta_l} \right) \underline{1} \\ &\leq \tilde{T}_{\delta_0} \tilde{T}_{\delta_1} \cdots \tilde{T}_{\delta_{k-1}} J + \epsilon\, \hat{\rho}_k\, \underline{1} \end{aligned}$$

where

$$\prod_{l=0}^{k-1} P_{\delta_l} = P_{\delta_0} P_{\delta_1} \cdots P_{\delta_{k-1}}$$

Similarly

$$\begin{aligned} \tilde{T}_{\delta_0} \tilde{T}_{\delta_1} \cdots \tilde{T}_{\delta_{k-1}} (J - \epsilon\, \underline{1}) &= \tilde{T}_{\delta_0} \tilde{T}_{\delta_1} \cdots \tilde{T}_{\delta_{k-1}} J - \epsilon \left(\prod_{l=0}^{k-1} P_{\delta_l} \right) \underline{1} \\ &\geq \tilde{T}_{\delta_0} \tilde{T}_{\delta_1} \cdots \tilde{T}_{\delta_{k-1}} J - \epsilon\, \hat{\rho}_k\, \underline{1} \end{aligned}$$

Let $J, J' \in \mathbf{R}^n$ and $\epsilon \geq 0$ be such that $\| J - J' \| \leq \epsilon$, where $\| \cdot \|$ is the sup norm defined by

$$\| J \| = \max_{i \in \{1,2,\ldots,n\}} |J(i)| \qquad \text{for } J \in \mathbf{R}^n$$

Thus

$$J' - \epsilon\, \underline{1} \leq J \leq J' + \epsilon\, \underline{1}$$

For $k \geq 1$, k integer and $\delta_0, \delta_1, \ldots, \delta_{k-1} \in \Lambda$, we have

$$\begin{aligned} \tilde{T}_{\delta_0}\tilde{T}_{\delta_1}\cdots\tilde{T}_{\delta_{k-1}}J' - \epsilon\,\hat{\rho}_k\,\underline{1} &\leq \tilde{T}_{\delta_0}\tilde{T}_{\delta_1}\cdots\tilde{T}_{\delta_{k-1}}J \\ &\leq \tilde{T}_{\delta_0}\tilde{T}_{\delta_1}\cdots\tilde{T}_{\delta_{k-1}}J' + \epsilon\,\hat{\rho}_k\,\underline{1} \end{aligned}$$

i.e.

$$\| \tilde{T}_{\delta_0}\tilde{T}_{\delta_1}\cdots\tilde{T}_{\delta_{k-1}}J - \tilde{T}_{\delta_0}\tilde{T}_{\delta_1}\cdots\tilde{T}_{\delta_{k-1}}J' \| \leq \epsilon\,\hat{\rho}_k$$

Let $k \in \mathbf{N}_0$. For $\mu \in \Upsilon$ let $\tilde{T}_\mu^k$ denote the composition of $\tilde{T}_\mu$ with itself k times; i.e. for $l \geq 1$, $\tilde{T}_\mu^l = \tilde{T}_\mu\tilde{T}_\mu^{l-1}$ with $\tilde{T}_\mu^0$ being the identity operator. Similarly for $\delta \in \Lambda$ let $\tilde{T}_\delta^k$ denote the composition of $\tilde{T}_\delta$ with itself k times, with $\tilde{T}_\delta^0$ being the identity operator. Likewise define $\tilde{T}^k$ to be the composition of $\tilde{T}$ with itself k times; i.e. for $l \geq 1$, $\tilde{T}^l = \tilde{T}\tilde{T}^{l-1}$, with $\tilde{T}^0$ being the identity operator.

Let $k \geq 1$, k integer, $J \in \mathbf{R}^n$ and scalar $\epsilon \geq 0$. For $\delta \in \Lambda$,

$$\begin{aligned} \tilde{T}_\delta^k\,(J + \epsilon\underline{1}) &\leq \tilde{T}_\delta^k J + \epsilon\,\rho_{\delta,k}\,\underline{1} \\ \tilde{T}_\delta^k\,(J - \epsilon\underline{1}) &\geq \tilde{T}_\delta^k J - \epsilon\,\rho_{\delta,k}\,\underline{1} \end{aligned}$$

For $J, J' \in \mathbf{R}^n$

$$\| \tilde{T}_\delta^k J - \tilde{T}_\delta^k J' \| \leq \rho_{\delta,k} \| J - J' \|$$

Similarly for $\mu \in \Upsilon$

$$\| \tilde{T}_\mu^k J - \tilde{T}_\mu^k J' \| \leq \rho_{\mu,k} \| J - J' \|$$

Let $J \in \mathbf{R}^n$. Fix scalar $\epsilon \geq 0$ and integer $k \geq 1$. Let $\mu_l \in \Upsilon$ for $l = 0, 1, \ldots, k-1$ be such that

$$\tilde{T}_{\mu_{k-1}}J = \tilde{T}J$$

$$\begin{aligned}
\tilde{T}_{\mu_{k-2}}\tilde{T}J &= \tilde{T}^2 J \\
&\vdots \\
\tilde{T}_{\mu_0}\tilde{T}^{k-1}J &= \tilde{T}^k J
\end{aligned}$$

i.e.

$$\tilde{T}_{\mu_{k-l}}\tilde{T}^{l-1}J = \tilde{T}^l J \qquad \text{for } l = 1, 2, \ldots, k$$

Hence

$$\begin{aligned}
\tilde{T}^k\,(J + \epsilon\,\underline{1}) &\leq \tilde{T}_{\mu_0}\tilde{T}_{\mu_1}\cdots\tilde{T}_{\mu_{k-1}}\,(J + \epsilon\,\underline{1}) \\
&= \tilde{T}^k J + \epsilon\left(\prod_{l=0}^{k-1} P_{\mu_l}\right)\underline{1} \\
&\leq \tilde{T}^k J + \epsilon\,\hat{\rho}_k\,\underline{1}
\end{aligned}$$

Here

$$\prod_{l=0}^{k-1} P_{\mu_l} = P_{\mu_0}P_{\mu_1}\cdots P_{\mu_{k-1}}$$

Let $\tilde{\mu}_l \in \Upsilon$ for $l = 0, 1, \ldots, k-1$ be such that

$$\begin{aligned}
\tilde{T}_{\tilde{\mu}_{k-1}}\,(J - \epsilon\,\underline{1}) &= \tilde{T}\,(J - \epsilon\,\underline{1}) \\
\tilde{T}_{\tilde{\mu}_{k-2}}\tilde{T}\,(J - \epsilon\,\underline{1}) &= \tilde{T}^2\,(J - \epsilon\,\underline{1}) \\
&\vdots \\
\tilde{T}_{\tilde{\mu}_0}\tilde{T}^{k-1}\,(J - \epsilon\,\underline{1}) &= \tilde{T}^k\,(J - \epsilon\,\underline{1})
\end{aligned}$$

i.e.

$$\tilde{T}_{\tilde{\mu}_{k-l}}\tilde{T}^{l-1}\,(J - \epsilon\,\underline{1}) = \tilde{T}^l\,(J - \epsilon\,\underline{1}) \qquad \text{for } l = 1, 2, \ldots, k$$

Now

$$\tilde{T}^k\,(J - \epsilon\,\underline{1}) = \tilde{T}_{\tilde{\mu}_0}\tilde{T}_{\tilde{\mu}_1}\cdots\tilde{T}_{\tilde{\mu}_{k-1}}\,(J - \epsilon\,\underline{1})$$

$$
\begin{aligned}
&= \tilde{T}_{\tilde{\mu}_0}\tilde{T}_{\tilde{\mu}_1}\cdots\tilde{T}_{\tilde{\mu}_{k-1}}J - \epsilon\left(\prod_{l=0}^{k-1}P_{\tilde{\mu}_l}\right)\underline{1}\\
&\geq \tilde{T}^k J - \epsilon\left(\prod_{l=0}^{k-1}P_{\tilde{\mu}_l}\right)\underline{1}\\
&\geq \tilde{T}^k J - \epsilon\,\hat{\rho}_k\,\underline{1}
\end{aligned}
$$

Thus for $J, J' \in \mathbf{R}^n$, if $\| J - J' \| \leq \epsilon$ for some scalar $\epsilon \geq 0$, then

$$J' - \epsilon\,\underline{1} \leq J \leq J' + \epsilon\,\underline{1}$$

Hence

$$\tilde{T}^k J' - \epsilon\,\hat{\rho}_k\,\underline{1} \leq \tilde{T}^k\,(J' - \epsilon\,\underline{1}) \leq \tilde{T}^k J \leq \tilde{T}^k\,(J' + \epsilon\,\underline{1}) \leq \tilde{T}^k J' + \epsilon\,\hat{\rho}_k\,\underline{1}$$

Thus for integer $k \geq 1$, we have

$$\| \tilde{T}^k J - \tilde{T}^k J' \| \leq \epsilon\,\hat{\rho}_k$$

Note that

$$\| \tilde{T} J - \tilde{T} J' \| \leq \| J - J' \|$$

since $0 \leq \hat{\rho}_1 \leq 1$. Hence $\tilde{T}$ is a non-expansion. Similarly for $\delta \in \Lambda$

$$\| \tilde{T}_\delta J - \tilde{T}_\delta J' \| \leq \| J - J' \|$$

since $0 \leq \rho_{\delta,1} \leq 1$.

C.4.2 Error Bounds For SSP MDPs

In this subsection we give some variants of the error bounds for the SSP problem given in [11].

Lemma C.7 *Let $J \in \mathbf{R}^n$. For any stationary deterministic proper policy $\mu \in \Upsilon$, let $J' = \tilde{T}_\mu J$. Let $\bar{c} = \max_{i \in \{1,2,\ldots,n\}} (J'(i) - J(i))$. Then*

$$\tilde{J}^\mu - J \leq \tilde{J}^\mu - J' + \bar{c}\underline{1} \leq \bar{c}\bar{N}_\mu$$

where the inequality is component wise. Here $\tilde{J}^\mu$ denotes the expected cost to go function for the SSP *problem under policy μ (which is the unique vector satisfying $\tilde{T}_\mu \tilde{J}^\mu = \tilde{J}^\mu$) and $\bar{N}_\mu \in \mathbf{R}^n$ is such that $\bar{N}_\mu(i)$ is the expected number of steps required to reach the terminal state 0 starting from state $i \in \{1, 2, \ldots, n\}$ under policy μ.*

□

Refer Chapter 1 and [11, 12] for more on SSP problems.

Proof of Lemma C.7

Now

$$\begin{aligned} J' &= \bar{g}^\mu + P_\mu J \\ \tilde{J}^\mu &= \bar{g}^\mu + P_\mu \tilde{J}^\mu \end{aligned}$$

Hence

$$\tilde{J}^\mu - J' \leq \tilde{J}^\mu - J' + \bar{c}\,\underline{1} = P_\mu \left(\tilde{J}^\mu - J\right) + \bar{c}\,\underline{1}$$

Multiplying this relation by P_μ and adding $\bar{c}\,\underline{1}$

$$P_\mu \left(\tilde{J}^\mu - J\right) + \bar{c}\,\underline{1} \leq P_\mu^2 \left(\tilde{J}^\mu - J\right) + \bar{c}\,(I + P_\mu)\,\underline{1}$$

Here I is the $n \times n$ identity matrix. Similarly continuing, we have for any integer $l \geq 1$

$$P_\mu^l \left(\tilde{J}^\mu - J\right) + \bar{c} \left(\sum_{k=0}^{l-1} P_\mu^k\right) \underline{1} \leq P_\mu^{l+1} \left(\tilde{J}^\mu - J\right) + \bar{c} \left(\sum_{k=0}^{l} P_\mu^k\right) \underline{1}$$

Here $P_\mu^0 = I$. Hence for $l \geq 1$

$$\tilde{J}^\mu - J \leq \tilde{J}^\mu - J' + \bar{c}\,\underline{1} \leq P_\mu^l \left(\tilde{J}^\mu - J\right) + \bar{c} \left(\sum_{k=0}^{l-1} P_\mu^k\right) \underline{1}$$

Now since $\lim_{l\to\infty} P_\mu^l$ is the zero matrix (μ being proper)

$$\lim_{l\to\infty} \left(\sum_{k=0}^{l-1} P_\mu^k\right) \underline{1} = (I - P_\mu)^{-1}\,\underline{1} = \bar{N}_\mu$$

Thus

$$\tilde{J}^\mu - J \leq \tilde{J}^\mu - J' + \bar{c}\,\underline{1} \leq \bar{c}\,\bar{N}_\mu$$

□

For the following lemma assume all stationary deterministic policies are proper.The lemma is also valid under the classical assumption, namely

Assumption C.1 *Let the following hold.*

1. *There exists at least one stationary deterministic policy*

2. *For every improper stationary deterministic policy μ, the corresponding cost to go $\tilde{J}^\mu(i)$ is ∞ for at least one state $i \in \{1, 2, \ldots, n\}$; i.e. some component of the sum $\sum_{k=0}^{N-1} P_\mu^k \bar{g}^\mu$ diverges to ∞ as $N \to \infty$.*

□

Lemma C.8 *Suppose $J \in \mathbf{R}^n$. Let $J' \in \mathbf{R}^n$ be such that $J' = \tilde{T}J$. If $\underline{c} = \min_{i\in\{1,2,\ldots,n\}} (J'(i) - J(i))$, then*

$$\underline{c}\bar{N}_{\mu^*} \leq \tilde{J}^* - J' + \underline{c}\underline{1} \leq \tilde{J}^* - J$$

Here μ^* is any optimal stationary deterministic policy (which is proper) and $\tilde{J}^*$ is the optimal cost to go vector for the SSP problem.

□

Note that $\tilde{T}J = \min_{\mu \in \Upsilon} \tilde{T}_\mu J$, where the minimization is taken component wise. J^* is the unique vector which satisfies the Bellman equation $\tilde{T}\tilde{J}^* = \tilde{J}^*$.

Proof of Lemma C.8

Now (see also [11])

$$\begin{aligned} J' &\leq \bar{g}^{\mu^*} + P_{\mu^*} J \\ \tilde{J}^* &= \bar{g}^{\mu^*} + P_{\mu^*} \tilde{J}^* \end{aligned}$$

Hence

$$\tilde{J}^* - J \geq \tilde{J}^* - J' + \underline{c}\,\underline{1} \geq P_{\mu^*}\left(\tilde{J}^* - J\right) + \underline{c}\,\underline{1}$$

Multiplying this relation by P_{μ^*} and adding $\underline{c}\,\underline{1}$

$$P_{\mu^*}\left(\tilde{J}^* - J\right) + \underline{c}\,\underline{1} \geq P_{\mu^*}^2\left(\tilde{J}^* - J\right) + \underline{c}\,(I + P_{\mu^*})\,\underline{1}$$

Similarly continuing, we have for any integer $l \geq 1$

$$P_{\mu^*}^l\left(\tilde{J}^* - J\right) + \underline{c}\left(\sum_{k=0}^{l-1} P_{\mu^*}^k\right)\underline{1} \geq P_{\mu^*}^{l+1}\left(\tilde{J}^* - J\right) + \underline{c}\left(\sum_{k=0}^{l} P_{\mu^*}^k\right)\underline{1}$$

Hence for $l \geq 1$

$$\tilde{J}^* - J \geq \tilde{J}^* - J' + \underline{c}\underline{1} \geq P_{\mu^*}^l\left(\tilde{J}^* - J\right) + \underline{c}\left(\sum_{k=0}^{l-1} P_{\mu^*}^k\right)\underline{1}$$

Since μ^* is proper we have $\lim_{l\to\infty} P_{\mu^*}^l$ equals the $n \times n$ zero matrix. Hence

$$\lim_{l\to\infty}\left(\sum_{k=0}^{l-1} P_{\mu^*}^k\right)\underline{1} = (I - P_{\mu^*})^{-1}\,\underline{1} = \bar{N}_{\mu^*}$$

So we have

$$\tilde{J}^* - J \geq \tilde{J}^* - J' + \underline{c}\,\underline{1} \geq \underline{c}\,\bar{N}_{\mu^*}$$

□

The preceding lemma is also true for any proper stationary randomized policy δ^* which is optimal, i.e. δ^* such that $\tilde{T}_{\delta^*}\tilde{J}^* = \tilde{T}\tilde{J}^* = \tilde{J}^*$.

As a corollary we have

Corollary C.1 *Suppose $J \in \mathbf{R}^n$. Let μ be any stationary deterministic proper policy and $J' \in \mathbf{R}^n$ be such that $J' = \tilde{T}_\mu J$. If $\underline{c} = \min_{i\in\{1,2,\ldots,n\}} (J'(i) - J(i))$ then*

$$\underline{c}\,\bar{N}_\mu \leq \tilde{J}^\mu - J' + \underline{c}\,\underline{1} \leq \tilde{J}^\mu - J$$

□

Corollary C.1 and Lemma C.7 hold also for any proper stationary randomized policy.

C.4.3 Approximate Policy Iteration Bounds For SSP Problems

In this subsection we give error bounds for approximate policy iteration [12, 16]. Assume that all stationary deterministic policies are proper. Let $\{\mu_k\}$ be a sequence of stationary deterministic policies and $\{\tilde{J}_k\}$ be a corresponding sequence of approximate cost vectors satisying

$$\| \tilde{J}_k - \tilde{J}^{\mu_k} \| \leq \epsilon \qquad \text{for } k = 0, 1, 2, \ldots$$

and

$$\| \tilde{T}\tilde{J}_k - \tilde{T}_{\mu_{k+1}}\tilde{J}_k \| \leq \varepsilon \qquad \text{for } k = 0, 1, 2, \ldots$$

μ_0 is chosen arbitrarily. ϵ and ε are non negative scalars. The above scheme is called an approximate policy iteration for the SSP problem.

Here $\tilde{J}^{\mu_k}$ is the cost to go vector corresponding to policy μ_k for $k = 0, 1, 2, \ldots$. Let integer $m \geq 1$ be such that $\tilde{\rho}_m < 1$. Note that $\tilde{\rho}_n < 1$ (since all stationary deterministic policies are proper) and $\tilde{\rho}_k \downarrow_k$. Let $\tilde{J}^*$ be the optimal cost to go vector for the SSP. Also let $\| \cdot \|$ denote the sup norm. Note that

$$\rho_{\mu,k} = \max_{i \in \{1,2,\ldots,n\}} \Pr^{\mu} [s_k \neq 0 \mid s_0 = i] \qquad \text{for } \mu \in \Upsilon,\ k \in \mathrm{N}_0$$

and

$$\tilde{\rho}_k = \max_{\mu \in \Upsilon} \rho_{\mu,k} \qquad \text{for } k \in \mathbf{N}_0$$

Lemma C.9 *Assume that the stationary deterministic proper policies μ_k are generated by the approximate policy iteration. Then (see [12, 16])*

$$\limsup_{k \to \infty} \| \tilde{J}^{\mu_k} - \tilde{J}^* \| \leq \frac{m(1 - \tilde{\rho}_m + m)(\varepsilon + 2\epsilon)}{(1 - \tilde{\rho}_m)^2}$$

□

The above result also holds when the SSP problem satisfy Assumption C.1 and all the stationary deterministic policies μ_ks generated by the approximate policy iteration are proper. In this case we redefine $\tilde{\rho}_k$ (just for this case) as

$$\tilde{\rho}_k = \max_{\mu \in \Upsilon,\ \mu \text{ proper}} \rho_{\mu,k}$$

for $k \in \mathbf{N}_0$.

C.4.4 Some Observations On SSP MDPs

A stationary randomized policy $\delta \in \Lambda$ is proper if and only if the Markov chain (with states $\mathcal{S} = \{0, 1, 2, \ldots, n\}$) corresponding to policy δ has only one recurrent class, namely $\{0\}$.

Let $\delta \in \Lambda$, We say that δ *subsumes* a stationary deterministic policy $\mu \in \Upsilon$ if and only if $[\delta(i)]_{\mu(i)} > 0$ for $i \in \{1, 2, \ldots, n\}$; i.e. action $\mu(i)$ is taken with positive probability from state i under policy δ for $i \in \{1, 2, \ldots, n\}$.

Now it can be seen that a stationary randomized policy δ is proper if and only if there exists a proper stationary deterministic policy $\mu \in \Upsilon$ subsumed by δ. The if part is easy to prove. The only if part can be proven using Theorem B.1 of Appendix B.

The proof of the following proposition is given in [12], but we give it here since it is short and illustrative.

Proposition C.3 *Under Assumption C.1 it can be shown that a stationary deterministic policy $\mu \in \Upsilon$ satisfying for some vector $J \in \mathbf{R}^n$, $J \geq \tilde{T}_\mu J$ (i.e. $J(i) \geq \left(\tilde{T}_\mu J\right)(i),\ i = 1, 2, \ldots, n$) is proper.*

□

Proof of Proposition C.3

Let $J \in \mathbf{R}^n$ be such that $J \geq \tilde{T}_\mu J$. Now it is easy to see that for integer $k \geq 1$

$$\tilde{T}_\mu^k J = P_\mu^k J + \sum_{l=0}^{k-1} P_\mu^l \bar{g}^\mu$$

Hence by monotonicity of $\tilde{T}_\mu$, we have for integer $k \geq 1$

$$J \geq \tilde{T}_\mu^k J = P_\mu^k J + \sum_{l=0}^{k-1} P_\mu^k \bar{g}^\mu$$

Since P_μ^k is a substochastic matrix, the components of $P_\mu^k J$ is bounded. If μ were not proper, by Assumption C.1 part 2, some component of the sum in the right-hand side of the above relation would diverge to ∞ as $k \to \infty$, which is a contradiction.

□

We may extend the above result to stationary randomized policies too.

Proposition C.4 *Under Assumption C.1 it can be shown that a stationary randomized policy $\delta \in \Lambda$ satisfying for some vector $J \in \mathbf{R}^n$, $J \geq \tilde{T}_\delta J$ (i.e. $J(i) \geq \left(\tilde{T}_\delta J\right)(i),\ i = 1, 2, \ldots, n$) is proper.*

□

Proof of Proposition C.4

Suppose for $J \in \mathbf{R}^n$ we have $\tilde{T}_\delta J \leq J$. That is for each $i \in \{1, 2, \ldots, n\}$

$$\sum_{a \in \mathcal{A}(i)} [\delta(i)]_a \left(g(i,a) + \sum_{j=1}^n p_{ij}(a) J(j) \right) \leq J(i)$$

This implies that for each $i \in \{1, 2, \ldots, n\}$, there exists $\tilde{a}_i \in \mathcal{A}(i)$ such that

$$\left(g(i, \tilde{a}_i) + \sum_{j=1}^n p_{ij}(\tilde{a}_i) J(j) \right) \leq J(i)$$

and $[\delta(i)]_{\tilde{a}_i} > 0$. Let $\mu(i) = \tilde{a}_i$ for $i \in \{1, 2, \ldots, n\}$. Then

$$\tilde{T}_\mu J \leq J$$

implying that the stationary deterministic policy μ is proper. Since δ subsumes μ, δ is proper.

□

Proposition C.5 *Under Assumption C.1 it can be shown that for every improper stationary randomized policy δ, the corresponding cost to go $\tilde{J}^{\delta}(i)$ is ∞ for at least one state $i \in \{1, 2, \ldots, n\}$; i.e. some component of the sum $\sum_{k=0}^{N-1} P_{\delta}^{k} \bar{g}^{\delta}$ diverges to ∞ as $N \to \infty$.*

□

Proof of Proposition C.5

Let δ be an improper stationary randomized policy. Let $J_0 = \underline{0}$, where $\underline{0} \in \mathbf{R}^n$ is the zero vector. Let $J_k = \tilde{T}_{\delta}^{k} J_0 = \sum_{l=0}^{k-1} P_{\delta}^{l} \bar{g}^{\delta}$. Now $\liminf_{k \to \infty} J_k(i) \geq \tilde{J}^{*}(i)$ for $i \in \{1, 2, \ldots, n\}$. Here $\tilde{J}^{*} \in \mathbf{R}^n$ is the optimal cost to go vector for the SSP problem.

Let $J(i) = \liminf_{k \to \infty} J_k(i)$ for $i \in \{1, 2, \ldots, m\}$. $J(i)$ is bounded below by $\tilde{J}^{*}(i)$ for $i \in \{1, 2, \ldots, n\}$. We have to prove that $J(i)$ is ∞ for at least one $i \in \{1, 2, \ldots, n\}$. Suppose not. That is $J \in \mathbf{R}^n$. Given any scalar $\epsilon > 0$, there exists a positive integer N_{ϵ} such that $J_l \geq J - \epsilon \, \underline{1}$ for $l \geq N_{\epsilon}$. Here $\underline{1} \in \mathbf{R}^n$ is the vector with all components equal to one. Hence for $l \geq N_{\epsilon}$ we have

$$\tilde{T}_{\delta} J_l \geq \tilde{T}_{\delta} J - \epsilon \, P_{\delta} \, \underline{1} \geq \tilde{T}_{\delta} J - \epsilon \, \underline{1}$$

Fix an $i \in \{1, 2, \ldots, n\}$. Now there exists $k \geq N_{\epsilon}$ such that $J_{k+1}(i) \leq J(i) + \epsilon$. That is

$$\left(\tilde{T}_{\delta} J_k \right)(i) \leq J(i) + \epsilon$$

Hence

$$J(i) + \epsilon \geq \left(\tilde{T}_\delta J\right)(i) - \epsilon$$

That is

$$2\,\epsilon + J(i) \geq \left(\tilde{T}_\delta J\right)(i)$$

Since this is true for any $\epsilon > 0$ we have $J(i) \geq \left(\tilde{T}_\delta J\right)(i)$. Thus we have

$$\tilde{T}_\delta J \leq J$$

Hence by Proposition C.4, δ is proper; a contradiction.

□

Let $\| \cdot \|$ be the sup norm on $\mathbf{R}^n$. Let $J, J' \in \mathbf{R}^n$. For $\delta \in \Lambda$ and integer $m \geq 1$

$$\| \tilde{T}_\delta^m J - \tilde{T}_\delta^m J' \| \leq \rho_{\delta,m} \| J - J' \|$$

For proper stationary randomized policy δ, we have $0 \leq \rho_{\delta,n} < 1$. Also $\rho_{\delta,k} \downarrow_k$. Let $0 \leq \rho_{\delta,m} < 1$. Then $\tilde{T}_\delta$ is an m-stage contraction mapping (and also a non-expansion) with respect to the sup norm $\| \cdot \|$, and has a contraction coefficient $\rho_{\delta,m}$. Hence if for $J \in \mathbf{R}^n$,

$$\| \tilde{T}_\delta J - J \| \leq \epsilon$$

for some scalar $\epsilon \geq 0$, then the cost to go vector (function) for the SSP problem under policy δ, namely $\tilde{J}^\delta$ satisfies

$$\| \tilde{J}^\delta - J \| \leq \frac{m\epsilon}{1 - \rho_{\delta,m}}$$

by Proposition C.2.

Similarly for integer $m \geq 1$ and $J, J' \in \mathbf{R}^n$, we have

$$\| \tilde{T}^m J - \tilde{T}^m J' \| \leq \hat{\rho}_m \| J - J' \|$$

If all stationary deterministic policies are proper then $0 \leq \hat{\rho}_n < 1$ and also $\hat{\rho}_k \downarrow_k$. Let $0 \leq \hat{\rho}_m < 1$. Then $\tilde{T}$ is an m-stage contraction mapping (and also a non-expansion) with respect to the sup norm $\| \cdot \|$ and has a contraction coefficient $\hat{\rho}_m$. Hence if for $J \in \mathbf{R}^n$,

$$\| \tilde{T} J - J \| \leq \epsilon$$

for some scalar $\epsilon \geq 0$, then by Proposition C.2

$$\| \tilde{J}^* - J \| \leq \frac{m\epsilon}{1 - \hat{\rho}_m}$$

where $\tilde{J}^*$ is the optimal cost to go vector for the all stationary deterministic policies proper, SSP problem.

Consider an all stationary deterministic policies proper SSP problem. Let $0 \leq \hat{\rho}_m < 1$ for some integer $m \geq 1$. Consider the approximate value iteration scheme where we generate the sequence $\{J_k\}$ according to the restriction

$$\| J_{k+1} - \tilde{T} J_k \| \leq \epsilon$$

starting from some $J_0 \in \mathbf{R}^n$. Here ϵ is non-negative scalar. Then by Lemma C.1

$$\limsup_{k \to \infty} \| J_k - \tilde{J}^* \| \leq \frac{m\epsilon}{1 - \hat{\rho}_m}$$

where $\tilde{J}^*$ is the optimal cost to go vector for the SSP problem.

Consider an SSP problem with all stationary deterministic policies proper. Let $0 \leq \hat{\rho}_m < 1$ for some integer $m > 1$. Then $\tilde{T}$ is an m-stage contraction mapping

(also a non-expansion) under the sup norm $\| \cdot \|$ with contraction coefficient $\hat{\rho}_m$. Let $\tilde{J}^* \in \mathbf{R}^n$ be the optimal cost to go vector which is the unique fixed point of $\tilde{T}$. Let $J \in \mathbf{R}^n$ be such that

$$\| \tilde{J}^* - J \| \leq \epsilon$$

Let $\delta \in \Lambda$, which is also proper and an m-stage contraction (and non-expansion) be such that

$$\| \tilde{T}J - \tilde{T}_\delta J \| \leq \varepsilon$$

Here ϵ and ε are non-negative scalars. Then by Lemma C.3

$$\| \tilde{J}^* - \tilde{J}^\delta \| \leq \frac{(2(m-1) + (1+\hat{\rho}_m))\epsilon + m\varepsilon}{1 - \hat{\rho}_m} \tag{C.8}$$

where $\tilde{J}^\delta$ is the cost to go vector for stationary randomized policy δ. In the above relation C.8, $\hat{\rho}_m$ may be replaced by $\rho_{\delta,m}$.

For the above problem if $\mu \in \Upsilon$, which is also proper and an m-stage contraction (and a non-expansion) is such that

$$\| \tilde{T}J - \tilde{T}_\mu J \| \leq \varepsilon$$

then

$$\| \tilde{J}^* - \tilde{J}^\mu \| \leq \frac{(2(m-1) + (1+\tilde{\rho}_m))\epsilon + m\varepsilon}{1 - \tilde{\rho}_m} \tag{C.9}$$

where $\tilde{J}^\mu$ is the cost to go vector for the stationary deterministic policy μ. In the above relation C.9, $\tilde{\rho}_m$ may be replaced by $\rho_{\mu,m}$.

Let

$$\varsigma = \min_{\tilde{\mu} \in \Upsilon,\ \tilde{\mu} \text{ not optimal}} \| \tilde{J}^{\tilde{\mu}} - \tilde{J}^* \|$$

If not all stationary deterministic policies are optimal, then $\varsigma > 0$. For all sufficiently small ϵ and ε, we have $\tilde{J}^{\mu} = \tilde{J}^{*}$. This can be seen, since for sufficiently small ϵ and ε, the right hand side of the relation C.9 is less than ς.

Suppose $\hat{\rho}_1 < 1$ (note that $\tilde{\rho}_1 = \hat{\rho}_1$). Then $\tilde{T}$, $\tilde{T}_{\mu}$ and $\tilde{T}_{\delta}$ for $\mu \in \Upsilon,\ \delta \in \Lambda$ are one stage contraction mappings. For this case also the above bounds (relation C.8 and relation C.9) hold, but are looser than the bound $\frac{2\hat{\rho}_1\epsilon+\varepsilon}{1-\hat{\rho}_1}$ given by Lemma C.2.

Yet another observation is the following. Let $\delta \in \Lambda$ be a proper policy. For $J \in \mathbf{R}^n$,

$$\begin{aligned} \tilde{T}_{\delta}J - J &= \bar{g}^{\delta} + P_{\delta}J - J \\ &= \bar{g}^{\delta} - (I - P_{\delta})J \end{aligned}$$

where I is the $n \times n$ identity matrix. Pre-multiplying by $(I - P_{\delta})^{-1}$, we get

$$\begin{aligned} (I - P_{\delta})^{-1}(\tilde{T}_{\delta}J - J) &= (I - P_{\delta})^{-1}\bar{g}^{\delta} - J \\ &= \tilde{J}^{\delta} - J \end{aligned}$$

Here $\tilde{J}^{\delta} = (I - P_{\delta})^{-1}\bar{g}^{\delta}$ is the cost to go vector for the SSP problem under policy δ. Thus

$$\bar{N}_{\delta}(i) \min_{j\in\{1,2,\ldots,n\}} \Big((\tilde{T}_{\delta}J)(j) - J(j)\Big) \leq \tilde{J}^{\delta}(i) - J(i) \leq \bar{N}_{\delta}(i) \max_{j\in\{1,2,\ldots,n\}} \Big((\tilde{T}_{\delta}J)(i) - J(i)\Big)$$

for $i \in \{1, 2, \ldots, n\}$.

Here $\bar{N}_{\delta} \in \mathbf{R}^n$ is the vector with components $\bar{N}_{\delta}(i)$ equal to the expected number of stages to reach the terminal state 0, starting from state $i \in \{1, 2, \ldots, n\}$, under policy δ. i.e.

$$\bar{N}_{\delta}(i) = e_i^T (I - P_{\delta})^{-1}\underline{1}$$

Here $\underline{1} \in \mathbf{R}^n$ is the vector with all components equal to one and $e_i \in \mathbf{R}^n$ is the i^{th} coordinate vector whose i^{th} component is one and all other entries are zero. Note that $(I - P_\delta)^{-1} = \sum_{k=0}^{\infty} P_\delta^k$.

C.4.5 Weighted Sup-Norm Property Of "All Proper Policies" SSP MDP

For $J \in \mathbf{R}^n$ we define the weighted sup norm

$$\| J \|_\xi = \max_{i \in \{1,2,\ldots,n\}} \frac{|J(i)|}{\xi(i)}$$

where $\xi = (\xi(1), \xi(2), \ldots, \xi(n))^T \in \mathbf{R}^n$ has all components positive. In this subsection we assume that all stationary deterministic policies are proper.

We have the following result from [16, page 23]

Proposition C.6 *Suppose all stationary deterministic policies are proper for the* SSP *problem. Then there exists a vector $\xi \in \mathbf{R}^n$ with positive components, such that $\tilde{T}, \tilde{T}_\mu$ and $\tilde{T}_\delta$ for all $\mu \in \Upsilon$, $\delta \in \Lambda$ are contraction mappings with respect to the weighted sup norm $\| \cdot \|_\xi$. In particular there exists a contraction coefficient β, with $0 \leq \beta < 1$ such that*

$$\sum_{j=1}^{n} p_{ij}(u)\, \xi(j) \quad \leq \quad \beta\, \xi(i) \tag{C.10}$$

for $i \in \{1, 2, \ldots, n\}$ and $u \in \mathcal{A}(i)$.

□

Note that the above relation C.10 implies that given $\delta \in \Lambda$

$$\sum_{j=1}^{n} [P_\delta]_{ij}\, \xi(j) \quad \leq \quad \beta\, \xi(i) \tag{C.11}$$

for $i \in \{1, 2, \ldots, n\}$. Here

$$[P_\delta]_{ij} = \sum_{a \in \mathcal{A}(i)} [\delta(i)]_a\, p_{ij}(a)$$

$[\delta(i)]_a$ being the probability of taking action a from state i under policy δ.

The proof that $\tilde{T}$ and $\tilde{T}_\mu,\ \mu \in \Upsilon$, are contraction mappings follows [16] from the relation C.10. Extension to the case $\tilde{T}_\delta$ for $\delta \in \Lambda$, follows easily from the relation C.11.

One choice for ξ and β are as follows [16, page 24]. Consider a new SSP problem where the transition probabilities are the same as in the original SSP problem, but the immediate costs are equal to -1 corresponding to all feasible actions from all the states in $\{1, 2, \ldots, n\}$ (for the termination state 0, the self transition cost is zero). Let $\breve{J} \in \mathbf{R}^n$ be the optimal cost to go vector for the new problem. Then $\breve{J}$ satisfies the Bellman equation [11, 12]

$$\breve{J}(i) = -1 + \min_{u \in \mathcal{A}(i)} \sum_{j=1}^{n} p_{ij}(u) \breve{J}(j)$$

for $i \in \{1, 2, \ldots, n\}$. Define $\xi(i) = -\breve{J}(i)$ for $i = 1, 2, \ldots, n$. Then $\xi(i) \geq 1$ for $i \in \{1, 2, \ldots, n\}$ and

$$\sum_{j=1}^{n} p_{ij}(u)\, \xi(j) \leq \xi(i) - 1 \leq \beta\, \xi(i)$$

for $i = 1, 2, \ldots, n$ and $u \in \mathcal{A}(i)$. Here β is defined by

$$\beta = \max_{i \in \{1,2,\ldots,n\}} \frac{\xi(i) - 1}{\xi(i)} < 1$$

The above proposition says that given $J, J' \in \mathbf{R}^n$ and $\delta \in \Lambda$ we have

$$\| \tilde{T}_\delta J - \tilde{T}_\delta J' \|_\xi \leq \beta \| J - J' \|_\xi$$

and

$$\| \tilde{T} J - \tilde{T} J' \|_\xi \leq \beta \| J - J' \|_\xi$$

Note that if $J \in \mathbf{R}^n$ is such that $\| J - \tilde{T} J \|_\xi \leq \epsilon$ for some scalar $\epsilon \geq 0$, then $\| \tilde{J}^* - J \|_\xi \leq \frac{\epsilon}{1-\beta}$, where $\tilde{J}^*$ is the optimal cost to go vector for the original SSP problem. This follows easily from Proposition C.2.

Consider the approximate value iteration scheme in which we generate a sequence of vectors in $\mathbf{R}^n$, namely $\{J_k\}$ satisfying

$$\| J_{l+1} - \tilde{T} J_l \|_\xi \leq \epsilon$$

for some scalar $\epsilon \geq 0$, starting from an arbitrary vector $J_0 \in \mathbf{R}^n$. Then Lemma C.1 implies

$$\limsup_{l \to \infty} \| J_l - \tilde{J}^* \|_\xi \leq \frac{\epsilon}{1-\beta}$$

We can make the following observation too. Suppose $J \in \mathbf{R}^n$ is such that $\| J - \tilde{J}^* \|_\xi \leq \epsilon$ for some scalar $\epsilon \geq 0$. Let $\mu \in \Upsilon$ be a stationary deterministic policy such that $\| \tilde{T}_\mu J - \tilde{T} J \|_\xi \leq \varepsilon$ for some scalar $\varepsilon \geq 0$. Then by Lemma C.2 we have

$$\| \tilde{J}^\mu - \tilde{J}^* \|_\xi \leq \frac{2\,\beta\,\epsilon + \varepsilon}{1-\beta} \tag{C.12}$$

Here $\tilde{J}^\mu$ is the cost to go vector for policy μ. Suppose

$$\varsigma = \min_{\tilde{\mu} \in \Upsilon,\ \tilde{\mu} \text{ not optimal}} \| \tilde{J}^{\tilde{\mu}} - \tilde{J}^* \|_\xi$$

If all stationary deterministic policies are not optimal, then $\varsigma > 0$. For all sufficiently small ϵ and ε, we have $\tilde{J}^{\mu} = \tilde{J}^{*}$. This can be seen, since for sufficiently small ϵ and ε, the right hand side of the relation C.12 is less than ς.

Similarly if $\| J - \tilde{J}^{*} \|_{\xi} \leq \epsilon$ and $\delta \in \Lambda$ is such that $\| \tilde{T}J - \tilde{T}_{\delta}J \|_{\xi} \leq \varepsilon$, then

$$\| \tilde{J}^{\delta} - \tilde{J}^{*} \|_{\xi} \leq \frac{2\,\beta\,\epsilon + \varepsilon}{1 - \beta}$$

Here $\tilde{J}^{\delta} \in \mathbf{R}^{n}$ is the cost to go vector for policy δ.

C.5 Equivalent SSP Problem For Discounted Cost MDP

Consider the finite state, finite action homogeneous Discounted Cost Problem with state space $\mathcal{S} = \{1, 2, \ldots, n\}$ and control constraint sets $\mathcal{A}(i) = \{1, 2, \ldots, |\mathcal{A}(i)|\}$, for $i \in \mathcal{S}$. Let $\mathcal{A} = \cup_{i \in \mathcal{S}} \mathcal{A}(i)$ denote the action space. See Chapter 1 for more on notations. The state at time $t \in \mathbf{N}_0$ is denoted by $s_t \in \mathcal{S}$, the action taken at time t is denoted by $u_t \in \mathcal{A}(s_t)$. The immediate cost incurred at time t while taking action $u_t \in \mathcal{A}(s_t)$ from state s_t is denoted by $g_t \in \mathbf{R}$. For $i, j \in \mathcal{S},\ a \in \mathcal{A}(i)$, let $p_{ij}(a)$ denote $\Pr[s_{t+1} = j \mid s_t = i, u_t = a]$. We assume that the expected immediate costs are finite, i.e. $\mathrm{E}[|g_t| \mid s_t = i, u_t = a] < \infty$ for $i \in \mathcal{S}, a \in \mathcal{A}(i)$. Let $g(i, a, j) \equiv \mathrm{E}\left[g_t \mid s_t = i, u_t = a, s_{t+1} = j\right]$ for $i, j \in \mathcal{S},\ a \in \mathcal{A}(i)$. The expected immediate cost for taking action a from state i for $i \in \mathcal{S},\ a \in \mathcal{A}(i)$ is

$$\begin{aligned} g(i, a) &\equiv \mathrm{E}\left[g_t \mid s_t = i, u_t = a\right] \\ &= \sum_{j=1}^{n} p_{ij}(a)\, g(i, a, j) \end{aligned}$$

Let the discount factor be $\beta \in [0, 1)$. Let $\mathcal{P}^\nu(\cdot \mid i)$ denote the probability measure for the discounted cost problem given the admissible policy $\nu \in \mathcal{M}$ and initial state $s_0 = i$. See Chapter 1 for definition of the admissible policy. The state space Ω under consideration is the space of infinite sequences $(s_0, u_0, g_0, s_1, u_1, g_1, \ldots, s_t, u_t, g_t, \ldots)$ where $s_t \in \mathcal{S}, u_t \in \mathcal{A}(s_t), g_t \in \mathbf{R}$. Let $\mathrm{E}^\nu(\cdot \mid i)$ denote the corresponding expectation. The infinite horizon discounted cost under policy ν, starting from state $i \in \mathcal{S}$ is

$$\begin{aligned} J^\nu(i) &= \mathrm{E}^\nu \left[\sum_{t=0}^{\infty} \beta^t g_t \mid s_0 = i \right] \\ &= \lim_{k \to \infty} \mathrm{E}^\nu \left[\sum_{t=0}^{k-1} \beta^t g_t \mid s_0 = i \right] \end{aligned}$$

Consider an associated problem with additive cost, without discounting in which after choosing action u_t at time t we "toss a coin" independently of everything else and decide with probability β to continue or else with probability $1 - \beta$ decide to terminate at this stage (if it has not been already terminated before time t). Here $t \in \mathbf{N}_0$. If the termination occurs at time $\tilde{N}$ (random), the total additive cost is $\sum_{t=0}^{\tilde{N}} g_t$. We are interested in minimizing the expected value of this cost starting from each starting state $i \in \mathcal{S}$.

The probability that termination has not occured before time t is β^t.

This problem can be translated into the following equivalent homogeneous SSP problem with state space $\tilde{\mathcal{S}} = \{0, 1, 2, \ldots, n\}$ (i.e. with an additional termination state 0). For want of more notation (just in this section) we denote by $\tilde{s}_t$ the state at time $t \in \mathbf{N}_0$ for the equivalent SSP problem, $\tilde{u}_t$ the action taken at time t for the equiuvalent SSP problem, $\tilde{g}_t$ the immediate cost incurred at time t for the equivalent SSP problem. Here the control constraints are the same as in the original Discounted

Cost problem, with the control constraint for the terminal state being $\mathcal{A}(0) = \{1\}$. Let

$$\begin{aligned}
\tilde{p}_{ij}(a) &\equiv \Pr\left[\tilde{s}_{t+1} = j \mid \tilde{s}_t = i, \tilde{u}_t = a\right] \\
&= \beta\, p_{ij}(a) \qquad \text{for } i, j \in \{1, 2, \ldots, n\},\ a \in \mathcal{A}(i) \\
\tilde{p}_{i0} &\equiv \Pr\left[\tilde{s}_{t+1} = 0 \mid \tilde{s}_t = i, \tilde{u}_t = a\right] \\
&= 1 - \beta \qquad \text{for } i \in \{1, 2, \ldots, n\},\ a \in \mathcal{A}(i)
\end{aligned}$$

Also

$$\begin{aligned}
\tilde{p}_{00}(1) &\equiv \Pr\left[\tilde{s}_{t+1} = 0 \mid \tilde{s}_t = 0, \tilde{u}_t = 1\right] \\
&= 1
\end{aligned}$$

For $i, j \in \{1, 2, \ldots, n\},\ a \in \mathcal{A}(i)$

$$\Pr\left[\tilde{g}_t \in B \mid \tilde{s}_t = i, \tilde{u}_t = a, \tilde{s}_{t+1} = j\right] = \Pr\left[g_t \in B \mid s_t = i, u_t = a, s_{t+1} = j\right]$$

for B, Borel subset of $\mathbf{R}$. Here 'Pr' on the right hand side is for the original discounted problem and 'Pr' on the left hand side is for the equivalent SSP problem.

Also for $i \in \{1, 2, \ldots, n\},\ a \in \mathcal{A}(i)$ and Borel subset B of $\mathbf{R}$

$$\begin{aligned}
&\Pr\left[\tilde{g}_t \in B \mid \tilde{s}_t = i, \tilde{u}_t = a, \tilde{s}_{t+1} = 0\right] = \\
&= \Pr\left[g_t \in B \mid s_t = i, u_t = a\right] \\
&= \sum_{j=1}^{n} p_{ij}(a) \Pr\left[g_t \in B \mid s_t = i, u_t = a, s_{t+1} = j\right]
\end{aligned}$$

Here 'Pr' on the left hand side is for the equivalent SSP problem and 'Pr' on the two right hand side terms are for the original discounted cost problem. Also

$$\Pr\left[\{\tilde{g}_t = 0\} \mid \tilde{s}_t = 0, \tilde{u}_t = 1\right] = 1$$

Let $\tilde{\mathcal{P}}^{\nu}(\cdot \mid i)$ denote the probability measure for the equivalent SSP problem, given the admissible policy ν and an initial state $\tilde{s}_0 = i \in \tilde{\mathcal{S}}$. See Chapter 1 and Section C.2 for the definition of the admissible policy for the SSP problem. The state space under consideration, $\tilde{\Omega}$ is the space of infinite sequences $(\tilde{s}_0, \tilde{u}_0, \tilde{g}_0, \tilde{s}_1, \tilde{u}_1, \tilde{g}_1, \ldots, \tilde{s}_t, \tilde{u}_t, \tilde{g}_t, \ldots)$, where $\tilde{s}_t \in \tilde{\mathcal{S}},\ \tilde{u}_t \in \mathcal{A}$ and $\tilde{g}_t \in \mathbf{R}$. Let $\tilde{\mathrm{E}}^{\nu}(\cdot \mid i)$ denote the corresponding expectation.

Note that for the equivalent SSP problem

$$\begin{aligned} \mathrm{E}\left[\tilde{g}_t \mid \tilde{s}_t = i, \tilde{u}_t = a, \tilde{s}_{t+1} = j\right] &\equiv \tilde{g}(i,a,j) \\ &= g(i,a,j) \\ &\equiv \mathrm{E}\left[g_t \mid s_t = i, u_t = a, s_{t+1} = j\right] \end{aligned}$$

for $i, j \in \{1, 2, \ldots, n\}$ and $a \in \mathcal{A}(i)$. The expectation in the above relation on the left hand side is for the equivalent SSP problem while the expectation on the right hand side is for the original discounted problem. Also for $i \in \{1, 2, \ldots, n\}$ and $a \in \mathcal{A}(i)$

$$\begin{aligned} \mathrm{E}\left[\tilde{g}_t \mid \tilde{s}_t = i, \tilde{u}_t = a, \tilde{s}_{t+1} = 0\right] &\equiv \tilde{g}(i,a,0) \\ &= g(i,a) \\ &\equiv \mathrm{E}\left[g_t \mid s_t = i, u_t = a\right] \\ &= \sum_{j=1}^{n} p_{ij}(a)\, g(i,a,j) \end{aligned} \tag{C.13}$$

Note that for $i \in \{1, 2, \ldots, n\}$ and $a \in \mathcal{A}(i)$

$$\begin{aligned} \mathrm{E}\left[\tilde{g}_t \mid \tilde{s}_t = i, \tilde{u}_t = a\right] &\equiv \tilde{g}(i,a) \\ &= \sum_{j=1}^{n} \beta\, p_{ij}(a)\, \tilde{g}(i,a,j) + (1-\beta)\, \tilde{g}(i,a,0) \end{aligned}$$

$$
\begin{aligned}
&= \beta\, g(i,a) + (1-\beta)\, g(i,a) \\
&= g(i,a)
\end{aligned}
$$

Also

$$\tilde{g}(0,1) \equiv \mathrm{E}\left[\tilde{g}_t \mid \tilde{s}_t = 0, \tilde{u}_t = 1\right] = 0$$

Note that for the equivalent SSP problem, all stationary deterministic policies are proper. Also the expected additive cost for the SSP problem, starting from state $i \in \{1, 2, \ldots, n\}$ under policy ν is denoted by

$$
\begin{aligned}
\tilde{J}^{\nu}(i) &= \lim_{k\to\infty} \tilde{\mathrm{E}}^{\nu}\left[\sum_{t=0}^{k-1} \tilde{g}_t \mid \tilde{s}_0 = i\right] \\
&= \tilde{\mathrm{E}}^{\nu}\left[\sum_{t=0}^{\infty} \tilde{g}_t \mid \tilde{s}_0 = i\right]
\end{aligned}
$$

With slight abuse of notation we use ν to denote the admissible policy for the original Discounted Cost problem and also for the corresponding policy for the equivalent SSP problem with the only difference that if at time t, state $\tilde{s}_t = 0$, then the action chosen is $\tilde{u}_t = 1$ and the system remains at state 0 itself with probability one incurring zero cost, while for $\tilde{s}_t \in \{1, 2, \ldots, n\}$ the action choice is the same as in the original Discounted Cost problem.

For $i, j \in \{1, 2, \ldots, n\},\ a \in \mathcal{A}(j)$ and $t \in \mathbf{N}_0$

$$
\begin{aligned}
&\tilde{\mathcal{P}}^{\nu}\left[\tilde{s}_t = j, \tilde{u}_t = a \mid \tilde{s}_0 = i\right] \\
&= \tilde{\mathcal{P}}^{\nu}\left[\tilde{s}_t = j, \tilde{u}_t = a \mid \tilde{s}_0 = i, \boldsymbol{\mathcal{I}}_{[\tilde{s}_t \neq 0]} = 1\right] \tilde{\mathcal{P}}^{\nu}\left[\tilde{s}_t \neq 0 \mid \tilde{s}_0 = i\right] \\
&\quad + \underbrace{\tilde{\mathcal{P}}^{\nu}\left[\tilde{s}_t = j, \tilde{u}_t = a \mid \tilde{s}_0 = i, \boldsymbol{\mathcal{I}}_{[\tilde{s}_t \neq 0]} = 0\right]}_{0} \tilde{\mathcal{P}}^{\nu}\left[\tilde{s}_t = 0 \mid \tilde{s}_0\right] \\
&= \mathcal{P}^{\nu}\left[s_t = j, u_t = a \mid s_0 = i\right] \cdot \beta^t
\end{aligned}
$$

Here $\boldsymbol{\mathcal{I}}$ denotes the indicator function. Also

$$\tilde{\mathcal{P}}^{\nu}\left[\tilde{s}_t = 0, \tilde{u}_t = 1 \mid \tilde{s}_0 = i\right] = 1 - \beta^t$$

for $i \in \{1, 2, \ldots, n\}$.

Hence for $i \in \{1, 2, \ldots, n\}$

$$\begin{aligned}
&\tilde{\mathrm{E}}^{\nu}\left[\tilde{g}_t \mid \tilde{s}_0 = i\right] \\
&= \sum_{j=1}^{n} \sum_{a \in \mathcal{A}(j)} \tilde{\mathcal{P}}^{\nu}\left[\tilde{s}_t = j, \tilde{u}_t = a \mid \tilde{s}_0 = i\right] \tilde{g}(j, a) \\
&= \sum_{j=1}^{n} \sum_{a \in \mathcal{A}(j)} \beta^t \mathcal{P}^{\nu}\left[s_t = j, u_t = a \mid s_0 = i\right] g(j, a) \\
&= \mathrm{E}^{\nu}\left[\beta^t g_t \mid s_0 = i\right] \\
&= \beta^t \mathrm{E}^{\nu}\left[g_t \mid s_0 = i\right]
\end{aligned}$$

Hence for $i \in \{1, 2, \ldots, n\}$

$$\lim_{k \to \infty} \tilde{\mathrm{E}}^{\nu}\left[\sum_{t=0}^{k-1} \tilde{g}_t \mid \tilde{s}_0 = i\right] = \lim_{k \to \infty} \mathrm{E}^{\nu}\left[\sum_{t=0}^{k-1} \beta^t g_t \mid s_0 = i\right]$$

i.e. for $i \in \{1, 2, \ldots, n\}$

$$\tilde{J}^{\nu}(i) = J^{\nu}(i)$$

In particular for any Markov Randomized policy and stationary (randomized or deterministic) policy, the cost to go is the same for the original Discounted Cost problem and the equivalent SSP problem. Note also that value iteration produces identical iterates for the two equivalent problems.

C.5.1 Error Bounds For Discounted Cost MDPs

Consider the discounted cost problem with discount factor $\beta \in [0, 1)$. Let $J^{\nu}(i)$ denote the infinite horizon discounted cost for admissible policy $\nu \in \mathcal{M}$, starting

from state $i \in \mathcal{S} = \{1, 2, \ldots, n\}$. i.e.

$$J^\nu(i) = \lim_{k\to\infty} \mathrm{E}^\nu \left[\sum_{t=0}^{k-1} \beta^t g_t \mid s_0 = i \right]$$

Here g_t is the immediate cost at stage $t \in \mathbf{N}_0$. $J^\nu \in \mathbf{R}^n$, with $J^\nu(i)$ being its i^{th} component for $i \in \{1, 2, \ldots, n\}$, is called the cost to go vector corresponding to policy ν for the discounted cost problem. For stationary randomized policy $\delta \in \Lambda$, the cost to go vector is

$$J^\delta = (I - \beta P_\delta)^{-1} \bar{g}^\delta$$

where

$$\begin{aligned}
[P_\delta]_{ij} &= \sum_{a\in\mathcal{A}(i)} [\delta(i)]_a \, p_{ij}(a) \qquad \text{for } i, j \in \mathcal{S} \\
\bar{g}^\delta(i) &= \sum_{a\in\mathcal{A}(i)} [\delta(i)]_a \, g(i, a) \qquad \text{for } i \in \mathcal{S} \\
g(i, a) &= \mathrm{E}\left[g_t \mid s_t = i, u_t = a \right] \qquad \text{for } i \in \mathcal{S},\ a \in \mathcal{A}(i)
\end{aligned}$$

Here P_δ is the $n \times n$ transition probability matrix (a stochastic matrix) for the policy $\delta \in \Lambda$. $\bar{g}^\delta \in \mathbf{R}^n$ is the expected immediate cost vector for policy δ. $[\delta(i)]_a$ is probability of taking action $a \in \mathcal{A}(i)$ from state $i \in \mathcal{S}$.

In particular for stationary deterministic policy $\mu \in \Upsilon$, the cost to go vector is

$$J^\mu = (I - \beta P_\mu)^{-1} \bar{g}^\mu$$

For $\mu \in \Upsilon$ let the operator $T_\mu : \mathbf{R}^n \to \mathbf{R}^n$ be defined by $T_\mu J = \bar{g}^\mu + \beta P_\mu J$ for $J \in \mathbf{R}^n$. For $\delta \in \Lambda$ let the operator $T_\delta : \mathbf{R}^n \to \mathbf{R}^n$ be defined by $T_\delta J = \bar{g}^\delta + \beta P_\delta J$ for $J \in \mathbf{R}^n$. Let the operator $T : \mathbf{R}^n \to \mathbf{R}^n$ be defined by $TJ = \min_{\mu\in\Upsilon} T_\mu J$ for

$J \in \mathbf{R}^n$, the minimization is taken component wise. That is for $i \in \mathcal{S}$ and $J \in \mathbf{R}^n$

$$(TJ)(i) = \min_{a \in \mathcal{A}(i)} \left(g(i,a) + \beta \sum_{j=1}^{n} p_{ij}(a)\, J(j) \right)$$

T_μ, T_δ and T are contraction mappings under the sup norm with contraction coefficient β.

Let $J^* \in \mathbf{R}^n$ denote the optimal (or minimal) cost to go vector for the discounted cost problem, i.e. $J^*(i) = \inf_{\nu \in \mathcal{M}} J^\nu(i)$, for $i \in \mathcal{S}$. We have the following propositions which follow from the equivalent SSP problem for the Discounted Cost problem (see Lemma C.7, Lemma C.8). Note that the expected number of steps to reach the terminal state from any state $i \in \{1, 2, \ldots, n\}$ for the equivalent SSP problem is $\frac{1}{1-\beta}$ under any admissible policy.

Proposition C.7 *Let $J \in \mathbf{R}^n$ and $\mu \in \Upsilon$ be any stationary deterministic policy. Let $J' = T_\mu J$ and $\bar{c} = \max_{i \in \{1,2,\ldots,n\}} (J'(i) - J(i))$. Then*

$$J^\mu - J \leq J^\mu - J' + \bar{c}\, \underline{1} \leq \bar{c}\, \frac{1}{1-\beta}\, \underline{1}$$

□

Here the inequality is component wise and $\underline{1} \in \mathbf{R}^n$ is the vector with all components equal to one. J^μ is the cost to go vector for the discounted cost problem corresponding to stationary deterministic policy μ.

Proposition C.8 *Let $J \in \mathbf{R}^n$ and $J' \in \mathbf{R}^n$ be $J' = TJ$.*

Let $\underline{c} = \min_{i \in \{1,2,\ldots,n\}} (J'(i) - J(i))$. Then

$$\underline{c}\, \frac{1}{1-\beta}\, \underline{1} \leq J^* - J' + \underline{c}\, \underline{1} \leq J^* - J$$

□

Here J^* is the optimal cost to go vector for the discounted cost problem. The following corollary follows from Corollary C.1.

Corollary C.2 *Let $J \in \mathbf{R}^n$ and $\mu \in \Upsilon$ be a stationary deterministic policy. Let $J' = T_\mu J$ and $\underline{c} = \min_{i\in\{1,2,\ldots,n\}} (J'(i) - J(i))$. Then*

$$\underline{c}\,\frac{1}{1-\beta}\,\underline{1} \leq J^\mu - J' + \underline{c}\,\underline{1} \leq J^\mu - J$$

□

Proposition C.7 and Corollary C.2 hold also for stationary randomized policy δ.

We have the following lemma,

Lemma C.10 *Let $J \in \mathbf{R}^n$ and*

$$\begin{aligned} \bar{c} &= \max_{i\in\{1,2,\ldots,n\}} ((TJ)(i) - J(i)) \\ \underline{c} &= \min_{i\in\{1,2,\ldots,n\}} ((TJ)(i) - J(i)) \end{aligned}$$

Let $\mu \in \Upsilon$ be such that, $\| T_\mu J - TJ \| \leq \epsilon$, where $\epsilon \geq 0$ and $\| \cdot \|$ is the sup norm. That is

$$\max_{i\in\{1,2,\ldots,n\}} ((T_\mu J)(i) - (TJ)(i)) \leq \epsilon$$

Then

$$J^\mu(i) - J^*(i) \leq \frac{\beta}{1-\beta}(\bar{c} - \underline{c}) + \frac{\epsilon}{1-\beta}$$

for $i \in \{1, 2, \ldots, n\}$. □

Proof of Lemma C.10

Let

$$\tilde{c} = \max_{i \in \{1,2,\ldots,n\}} ((T_\mu J)(i) - J(i))$$

Then from Proposition C.8 and Proposition C.7 we have for $i \in \{1, 2, \ldots, n\}$

$$(TJ)(i) + \frac{\beta}{1-\beta}\underline{c} \leq J^*(i)$$

and

$$J^\mu(i) \leq (T_\mu J)(i) + \frac{\beta}{1-\beta}\tilde{c}$$

Thus

$$J^\mu(i) - J^*(i) \leq (T_\mu J)(i) - (TJ)(i) + \frac{\beta}{1-\beta}(\tilde{c} - \underline{c})$$

Now

$$\tilde{c} \leq \bar{c} + \epsilon$$

and

$$(T_\mu J)(i) - (TJ)(i) \leq \epsilon$$

From this the result follows.

□

We have the following proposition, which follows from Lemma C.2. But we give an alternate proof here. Here $\| \cdot \|$ is the sup norm.

Proposition C.9 *Consider an infinite horizon discounted cost problem with discount factor $\beta \in [0, 1)$. Let $J \in \mathbf{R}^n$ be such that $\| J - J^* \| \leq \epsilon$. Let $\mu \in \Upsilon$ be a stationary deterministic policy such that $\| T_\mu J - TJ \| \leq \varepsilon$. Here ϵ and ε are*

non-negative scalars and J^* is the optimal cost to go vector for the discounted cost problem. Then

$$\| J^\mu - J^* \| \leq \frac{2\,\beta\,\epsilon + \varepsilon}{1-\beta}$$

where J^μ is ther cost to go vector for policy μ.

□

Proof of Proposition C.9

Now T and T_μ are contraction mappings under the sup norm $\| \cdot \|$, with contraction coefficient β. Also $T_\mu J^\mu = J^\mu$ and $TJ^* = J^*$.

$$\begin{aligned}
\| J^\mu - J^* \| &\leq \| T_\mu J^\mu - J^* \| \\
&\leq \| T_\mu J^\mu - T_\mu J \| + \| T_\mu J - J^* \| \\
&\leq \beta \| J^\mu - J \| + \| TJ - J^* \| + \| TJ - T_\mu J \| \\
&\leq \beta \| J^\mu - J^* \| + \beta \| J^* - J \| + \beta \| J - J^* \| + \varepsilon \\
&\leq \beta \| J^\mu - J^* \| + 2\,\beta\,\epsilon + \varepsilon
\end{aligned}$$

Hence

$$\| J^\mu - J^* \| \leq \frac{2\,\beta\,\epsilon + \varepsilon}{1-\beta}$$

□

Similar results also hold for stationary randomized policy $\delta \in \Lambda$.

Let

$$\varsigma = \min_{\tilde{\mu} \in \Upsilon,\ \tilde{\mu} \text{ not optimal}} \| J^{\tilde{\mu}} - J^* \|$$

If not all stationary deterministic policies are optimal, then $\varsigma > 0$ since there are

only a finite stationary deterministic policies. Hence in the proposition above for all sufficiently small ϵ and ε, $\frac{2\,\beta\,\epsilon+\varepsilon}{1-\beta} < \varsigma$ and hence $J^{\mu} = J^{*}$.

In the remaining portion of this subsection $\| \cdot \|$ denotes the sup norm.

Let $J \in \mathbf{R}^n$ and $\epsilon \geq 0$ be such that $\| TJ - J \| \leq \epsilon$. Then it follows from Proposition C.2, that $\| J - J^{*} \| \leq \frac{\epsilon}{1-\beta}$, where β is the discount factor. Similarly let δ be a stationary randomized policy and $\| T_{\delta}J - J \| \leq \varepsilon$, where ε is a non-negative scalar. Then again by Proposition C.2, we have $\| J^{\delta} - J \| \leq \frac{\varepsilon}{1-\beta}$. Hence $\| J^{\delta} - J^{*} \| \leq \frac{\epsilon+\varepsilon}{1-\beta}$.

In fact if $TJ \leq J$ and $T_{\delta}J \leq J$, we have $TJ \leq T_{\delta}J \leq J$. $\| TJ - J \| \leq \epsilon$, $\| T_{\delta}J - J \| \leq \varepsilon$ and $0 \leq \varepsilon \leq \epsilon$. Hence

$$J - \frac{\epsilon}{1-\beta}\underline{1} \leq J^{*} \leq J$$

and

$$J - \frac{\varepsilon}{1-\beta}\underline{1} \leq J^{\delta} \leq J$$

Thus

$$\| J^{*} - J^{\delta} \| \leq \frac{\epsilon}{1-\beta}$$

Here $\underline{1} \in \mathbf{R}^n$ is the vector with all components equal to one.

Now consider the approximate value iteration scheme for the discounted cost problem. Starting with some $J_0 \in \mathbf{R}^n$ we have

$$\| J_{k+1} - TJ_k \| \leq \epsilon$$

for all $k \in \mathbf{N}_0$. Here ϵ is a non-negative scalar. Then from Lemma C.1,

$$\limsup_{k\to\infty} \| J_k - J^{*} \| \leq \frac{\epsilon}{1-\beta}$$

where J^* is the optimal cost to go vector for the discounted cost problem.

Also we have the following proposition.

Proposition C.10 *Let $\delta \in \Lambda$ be a stationary randomized policy and $J \in \mathbf{R}^n$. Then*

$$(I - \beta P_\delta)^{-1} (T_\delta J - J) + J = J^\delta$$

□

Proof of Proposition C.10

$$\begin{aligned} T_\delta J - J &= \bar{g}^\delta + \beta P_\delta J - J \\ &= \bar{g}^\delta - (I - \beta P_\delta) J \end{aligned}$$

Hence

$$\begin{aligned} (I - \beta P_\delta)^{-1} (T_\delta J - J) &= (I - \beta P_\delta)^{-1} - J \\ &= J^\delta - J \end{aligned}$$

□

Let

$$\begin{aligned} \bar{c} &= \max_{i \in \{1,2,\ldots,n\}} ((T_\delta J)(i) - J(i)) \\ \underline{c} &= \min_{i \in \{1,2,\ldots,n\}} ((T_\delta J)(i) - J(i)) \end{aligned}$$

Then by the above proposition we have

$$\frac{1}{1-\beta}\, \underline{c}\, \underline{1} \leq J^\delta - J \leq \frac{1}{1-\beta}\, \bar{c}\, \underline{1}$$

where $\underline{1} \in \mathbf{R}^n$ is the vector with all components equal to one. Note that $(I - \beta P_\delta)^{-1} = \sum_{k=0}^{\infty} \beta^k P_\delta^k$.

C.5.2 Approximate Policy Iteration Bounds For Discounted Cost MDPs

As before, for $J \in \mathbf{R}^n$

$$\begin{aligned} T_\mu J &= \bar{g}^\mu + \beta P_\mu J \qquad \text{for } \mu \in \Upsilon \\ TJ &= \min_{\mu \in \Upsilon} T_\mu J \end{aligned}$$

where the minimization is taken component wise. Here β is the discount factor. We give the approximate policy iteration error bounds in the lemma below. For a proof see [12].

Lemma C.11 *Let $\{\mu_k\}$ be a sequence of stationary deterministic policies and $\{J_k\}$ be the corresponding sequence of approximate cost vectors satisfying*

$$\begin{aligned} \| J_k - J^{\mu_k} \| &\leq \epsilon \qquad \textit{for } k = 0, 1, 2, \ldots \\ \| TJ_k - T_{\mu_{k+1}} J_k \| &\leq \varepsilon \qquad \textit{for } k = 0, 1, 2, \ldots \end{aligned}$$

Then

$$\limsup_{k \to \infty} \| J_{\mu_k} - J^* \| \leq \frac{\varepsilon + 2\,\beta\,\epsilon}{(1-\beta)^2}$$

□

Here J^* is the optimal cost to go function for the discounted cost problem and J^{μ_k} is the cost to go vector for policy μ_k.

C.6 Error Bounds For Average Cost Problem

We are considering a finite state finite action MDP with state space $\mathcal{S} = \{1, 2, \ldots, n\}$. The control constraint sets are $\mathcal{A}(i) = \{1, 2, \ldots, |\mathcal{A}(i)|\}$ for state $i \in \mathcal{S}$. Here $|\cdot|$ represents the cardinality of the set. Let $\mathcal{A} = \cup_{i \in \mathcal{S}} \mathcal{A}(i)$. $p_{ij}(a) = \Pr[s_{t+1} = j \mid s_t = i, u_t = a]$ for $i, j \in \mathcal{S},\ a \in \mathcal{A}(i)$. Here $s_t \in \mathcal{S}$ is the state at time t and u_t is the action taken at time t from state s_t. $g_t \in \mathbf{R}$ denotes the immediate cost incurred at time t when action $u_t \in \mathcal{A}(s_t)$ is taken from state s_t. Let $g(i, a) = \mathrm{E}[g_t \mid s_t = i, u_t = a]$ for $i \in \mathcal{S},\ a \in \mathcal{A}(i)$. Let $\mathcal{M}$ denote the set of admissible policies. See Chapter 1 for more on notations. Let $\mathcal{P}^\nu(\cdot \mid i)$ denote the probability measure for policy ν and starting state $s_0 = i$. Let $\mathrm{E}^\nu(\cdot \mid i)$ denote the corresponding expectation. Here the state space Ω under consideration is the space of infinite sequences $(s_0, u_0, g_0, s_1, u_1, g_1, \ldots, s_t, u_t, g_t, \ldots)$ where $s_t \in \mathcal{S},\ u_t \in \mathcal{A}$ and $g_t \in \mathbf{R}$. For an admissible policy $\nu \in \mathcal{M}$, $\bar{\vartheta}^\nu \in \mathbf{R}^n$ denotes the average cost to go vector. $\bar{\vartheta}^\nu(i)$ denotes the expected average cost starting from state $i \in \mathcal{S}$. i.e.

$$\bar{\vartheta}^\nu(i) = \limsup_{k \to \infty} \frac{1}{k} \sum_{t=0}^{k-1} \mathrm{E}^\nu[g_t \mid s_0 = i]$$

For stationary policies the limit exist (i.e. we can replace the lim sup with lim in the above equation). We are interested in minimizing this expected average costs for all initial states $i \in \mathcal{S}$. It is known [12, 40] that there exists a stationary deterministic policy (for example a Blackwell optimal policy) which is optimal.

Υ denotes the set of stationary deterministic policies. Λ denote the set of stationary randomized policies. Refer Chapter 4 for more on stationary randomized policies.

For $\mu \in \Upsilon$ let the operator $\bar{T}_\mu : \mathbf{R}^n \to \mathbf{R}^n$ be defined by

$$\bar{T}_\mu J = \bar{g}^\mu + P_\mu J \qquad \text{for } J \in \mathbf{R}^n.$$

For $i \in \mathcal{S}$, $\bar{g}^\mu(i) = g(i, \mu(i))$ is the expected immediate cost for taking action $\mu(i)$ from state i. P_μ is the $n \times n$ transition probability matrix (a stochastic matrix) corresponding to μ and is given by $[P_\mu]_{ij} = p_{ij}(\mu(i))$ for $i, j \in \mathcal{S}$.

For $\delta \in \Lambda$ define the operator $\bar{T}_\delta : \mathbf{R}^n \to \mathbf{R}^n$ by

$$\bar{T}_\delta J = \bar{g}^\delta + P_\delta J \qquad \text{for } J \in \mathbf{R}^n.$$

Here $\bar{g}^\delta$ is the expected immmediate cost vector for policy δ and P_δ is the transition probability matrix for policy δ. i.e.

$$\bar{g}^\delta(i) = \sum_{a \in \mathcal{A}(i)} [\delta(i)]_a \, g(i, a)$$

for $i \in \mathcal{S}$ and

$$[P_\delta]_{ij} = \sum_{a \in \mathcal{A}(i)} [\delta(i)]_a \, p_{ij}(a)$$

for $i, j \in \mathcal{S}$. Here $[\delta(i)]_a$ denotes the probability of taking action a from state i, under policy δ.

Define the operator $\bar{T} : \mathbf{R}^n \to \mathbf{R}^n$ by

$$\bar{T} J = \min_{\mu \in \Upsilon} \bar{T}_\mu J \qquad \text{for } J \in \mathbf{R}^n$$

where the minimization is taken component wise., i.e.

$$\left(\bar{T} J\right)(i) = \min_{a \in \mathcal{A}(i)} \left(g(i, a) + \sum_{j=1}^{n} p_{ij}(a) \, J(j) \right)$$

for $i \in \mathcal{S}$, $J \in \mathbf{R}^n$.

The operators, $\bar{T}_\mu, \bar{T}_\delta$ and $\bar{T}$ are all monotone operators which are non-expansions under the sup norm.

In the following $\underline{1} \in \mathbf{R}^n$ denotes the vector with all components equal to one. We have the following lemma which is a variant of the one in [10, page 325].

Lemma C.12 *Let $J \in \mathbf{R}^n$ and scalar $\epsilon \geq 0$. Let $\mu \in \Upsilon$ be such that $\bar{T}_\mu J \leq \bar{T}J + \epsilon \underline{1}$. Then for $i \in \mathcal{S}$,*

$$\begin{aligned} \min_{j \in \mathcal{S}} \left(\left(\bar{T}J \right)(j) - J(j) \right) &\leq \bar{\vartheta}^*(i) \\ &\leq \bar{\vartheta}^\mu(i) \\ &\leq \max_{j \in \mathcal{S}} \left(\left(\bar{T}J \right)(j) - J(j) \right) + \epsilon \end{aligned}$$

where $\bar{\vartheta}^*(i)$ is the optimal (minimal) average cost to go from state i. The bounds hold regardless of whether $\bar{\vartheta}^*(i)$ is independent of the initial state i.

□

Proof of Lemma C.12

$$\begin{aligned} \bar{T}_\mu J &\leq \left(\bar{T}J - J \right) + J + \epsilon \, \underline{1} \\ \bar{T}_\mu^2 J &\leq \bar{T}_\mu J + P_\mu \left(\bar{T}J - J \right) + \epsilon \, \underline{1} \\ &\leq \left(\bar{T}J - J \right) + P_\mu \left(\bar{T}J - J \right) + J + 2\epsilon \, \underline{1} \end{aligned}$$

Continuing similarly,

$$\bar{T}_\mu^N J \leq \sum_{k=0}^{N-1} P_\mu^k \left(\bar{T}J - J \right) + J + N\epsilon \, \underline{1}$$

for $N = 1, 2, \ldots$. Hence

$$\bar{\vartheta}^\mu = \lim_{N \to \infty} \frac{1}{N} \bar{T}_\mu^N J \leq P_\mu^* \left(\bar{T}J - J \right) + \epsilon \, \underline{1}$$

where the limit is taken component wise. Here

$$P_\mu^* = \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} P_\mu^k$$

which exists [12]. Hence

$$\bar{\vartheta}^\mu(i) \leq \max_{j\in\mathcal{S}} \left(\left(\bar{T}J \right)(j) - J(j) \right) + \epsilon$$

Let $\nu = (\delta_0, \delta_1, \delta_2, \ldots)$ be any Markov randomized policy, where $\delta_k \in \Lambda$,

We have for any $\delta \in \Lambda$,

$$\begin{aligned} \bar{T}_\delta J &\geq \left(\bar{T}J - J \right) + J \\ &\geq \min_{j\in\mathcal{S}} \left(\left(\bar{T}J \right)(j) - J(j) \right) \underline{1} + J \end{aligned} \tag{C.14}$$

We have for $N \in \mathbf{N}$ (here $\mathbf{N}$ is the set of positive integers)

$$\bar{T}_{\delta_N} J \geq \min_{j\in\mathcal{S}} \left(\left(\bar{T}J \right)(j) - J(j) \right) \underline{1} + J$$

Applying $\bar{T}_{\delta_{N-1}}$ to both sides of the above inequality and using inequality C.14,

$$\begin{aligned} \bar{T}_{\delta_{N-1}} \bar{T}_{\delta_N} J &\geq \bar{T}_{\delta_{N-1}} J + \min_{j\in\mathcal{S}} \left(\left(\bar{T}J \right)(j) - J(j) \right) \underline{1} \\ &\geq J + 2 \min_{j\in\mathcal{S}} \left(\left(\bar{T}J \right)(j) - J(j) \right) \underline{1} \end{aligned}$$

Continuing similarly

$$\bar{T}_{\delta_0} \bar{T}_{\delta_1} \cdots \bar{T}_{\delta_N} J \geq J + (N+1) \min_{j\in\mathcal{S}} \left(\left(\bar{T}J \right)(j) - J(j) \right) \underline{1}$$

Hence

$$\frac{1}{N+1} \left(\bar{T}_{\delta_0} \bar{T}_{\delta_1} \cdots \bar{T}_{\delta_N} J \right)(i) \geq \min_{j\in\mathcal{S}} \left(\left(\bar{T}J \right)(j) - J(j) \right) + \frac{J(i)}{N+1}$$

Hence

$$\begin{aligned}\limsup_{N\to\infty} \frac{1}{N+1}\left(\bar{T}_{\delta_0}\bar{T}_{\delta_1}\cdots\bar{T}_{\delta_N}J\right)(i) \\ \geq \liminf_{N\to\infty} \frac{1}{N+1}\left(\bar{T}_{\delta_0}\bar{T}_{\delta_1}\cdots\bar{T}_{\delta_N}J\right)(i) \\ \geq \min_{j\in\mathcal{S}}\left(\left(\bar{T}J\right)(j) - J(j)\right)\end{aligned}$$

Hence

$$\bar{\vartheta}^{\nu}(i) \geq \min_{j\in\mathcal{S}}\left(\left(\bar{T}J\right)(j) - J(j)\right) \tag{C.15}$$

Since ν is an arbitrary Markov randomized policy

$$\bar{\vartheta}^{*}(i) \geq \min_{j\in\mathcal{S}}\left(\left(\bar{T}J\right)(j) - J(j)\right)$$

Note that we need to focus only on Markov randomized policies, since given any history dependent randomized policy and an initial state $i \in \mathcal{S}$, there exists a Markov randomized policy, such that both of them have the same additive cost starting from initial state i (see Chapter 1 and also [40, Chapter 5, Theorem 5.5.3]).

Another way to look at this is that there exists a Blackwell optimal policy $\mu^* \in \Upsilon$ (identify the policy μ^* with ν in the above inequality C.15) such that

$$\bar{\vartheta}^{*}(i) = \bar{\vartheta}^{\mu^*}(i) \geq \min_{j\in\mathcal{S}}\left(\left(\bar{T}J\right)(j) - J(j)\right)$$

□

The following corollary follows easily from the above lemma.

Corollary C.3 *Let $\mu \in \Upsilon$ and $J \in \mathbf{R}^n$. Then for $i \in \{1, 2, \ldots, n\}$,*

$$\min_{j\in\mathcal{S}}\left(\left(\bar{T}_{\mu}J\right)(j) - J(j)\right) \leq \bar{\vartheta}^{\mu}(i) \leq \max_{j\in\mathcal{S}}\left(\left(\bar{T}_{\mu}J\right)(j) - J(j)\right)$$ □

Similar results hold for $\delta \in \Lambda$.

For a variant of the approximate policy iteration and corresponding error bounds, for average cost MDP which is unichain and has a common recurrent state see [12, 16].

Appendix D

Temporal Difference Schemes For Stochastic Shortest Path Problems

In this appendix we are interested in developing an approximate policy iteration scheme for Stochastic Shortest Path (SSP) problems, where all the stationary deterministic policies are proper. We use Temporal Difference (TD) Schemes [16] for evaluating the (undiscounted) cost to go function for a proper policy.

For detailed notations and formulation of the (homogeneous) Stochastic Shortest Path problem see Chapter 1 and Appendix C. We assume the state space to be $\mathcal{S} = \{0, 1, 2, \ldots, n\}$, with 0 being the zero cost absorption (termination) state. Here n is a positive integer. The state of the system at time $t \in \mathbf{N}_0$ is denoted by s_t which is an element of $\mathcal{S}$. $\mathbf{N}_0$ denotes the set of non-negative integers. The action taken at time t is denoted by u_t, where $u_t \in \mathcal{A}(s_t)$. $\mathcal{A}(i) = \{1, 2, \ldots, |\mathcal{A}(i)|\}$ denotes the finite control constraint set for $i \in \mathcal{S}$, and $|\mathcal{A}(i)|$ denotes the cardinality of the constraint set $\mathcal{A}(i)$. The immediate cost incurred at time $t \in \mathbf{N}_0$, while taking action u_t from state s_t is denoted by g_t, with $u_t \in \mathcal{A}(s_t)$. We assume that the expected immediate costs have finite (hence bounded) second moments; i.e. $\mathrm{E}\left[g_t^2 \mid s_t = i, u_t = a\right] < \infty$ for $i \in \mathcal{S},\ a \in \mathcal{A}(i)$. We assume that $\mathcal{A}(0) = \{1\}$ and the immediate cost incurred while taking action 1 from state 0 is zero, with the system remaining in state 0 with probability one. Since we assume all stationary deterministic policies to be proper, we also have that all stationary (randomized) policies are proper. For

a general admissible policy $\nu \in \mathcal{M}$, the expected infinite horizon non-discounted additive cost (or cost to go) starting from state $i \in \{1, 2, \ldots, n\}$, is defined by $\tilde{J}^{\nu}(i) = \limsup_{k\to\infty} \mathrm{E}^{\nu}\left[\sum_{t=0}^{k-1} g_t \mid s_0 = i\right]$. For the definition of the set of admissible or feasible policies $\mathcal{M}$, see Chapter 1 and Appendix C. Here $\mathrm{E}^{\nu}[\cdot \mid s_0 = i]$ denotes the expectation under the probability distribution induced by policy ν, starting from state $i \in \{1, 2, \ldots, n\}$ at time 0. Since we assume all stationary deterministic policies to be proper, the limit exists instead of the lim sup in the definition of $\tilde{J}^{\nu}(i)$. In fact

$$\tilde{J}^{\nu}(i) = \lim_{k\to\infty} \mathrm{E}^{\nu}\left[\sum_{t=0}^{k-1} g_t \mid s_0 = i\right] = \mathrm{E}^{\nu}\left[\sum_{t=0}^{\infty} g_t \mid s_0 = i\right]$$

for $i \in \{1, 2, \ldots, n\}$. The optimal cost to go vector $\tilde{J}^* \in \mathbf{R}^n$, is given by $\tilde{J}^*(i) = \inf_{\nu\in\mathcal{M}} \tilde{J}^{\nu}(i)$.

Before we proceed we restate the notations related to the Stationary Randomized Policies (see Chapter 4) with slight modifications for the setting of this appendix.

D.1 Stationary Randomized Policies

Define for each positive integer k,

$$\Delta_k \equiv \{(p_1, p_2, \ldots, p_k) \mid p_l \geq 0, \sum_{l=1}^{k} p_l = 1\}$$

the $k-1$ dimensional unit simplex. A stationary randomized policy (or a stochastic control kernel to be precise) for the SSP problem can be specified as

$$\delta \in \Lambda$$

where

$$\Lambda \equiv \Delta_{|\mathcal{A}(1)|} \times \Delta_{|\mathcal{A}(2)|} \times \cdots \times \Delta_{|\mathcal{A}(n)|}$$

Here $|\mathcal{A}(i)|$ denotes the cardinality of the control constraint set $\mathcal{A}(i)$ for state $i \in \{1, 2, \ldots, n\}$. For each $i \in \{1, 2, \ldots, n\}$

$$\begin{aligned} \delta(i) &\in \Delta_{|\mathcal{A}(i)|} \qquad \text{denotes} \\ [\delta(i)]_a &= \Pr(u_t = a \mid s_t = i), \qquad a \in \mathcal{A}(i) \end{aligned}$$

the probability of taking action a from state $i \in \{1, 2, \ldots, n\}$. It is implicitly assumed that the action taken from the terminal state 0, is the unique action 1 (note that $\mathcal{A}(0) = \{1\}$) under which the system remains at state 0 incurring zero cost.

For a particular stationary randomized policy $\delta \in \Lambda$, we obtain a homogeneous Markov Chain with state space $\mathcal{S}$ and transition probability defined as follows. For states $i \in \{1, 2, \ldots, n\}$ and $j \in \mathcal{S}$ we have the transition probability given by

$$p_{ij}^{\delta} = \sum_{a \in \mathcal{A}(i)} [\delta(i)]_a \, p_{ij}(a)$$

For $i, j \in \mathcal{S}$, $p_{ij}(a)$ is the probability that the next state is j, given that the current state is i and the action taken from state i is a. Also, $p_{00}^{\delta} = 1$. The expected immediate cost from state $i \in \{1, 2, \ldots, n\}$ is given by

$$\bar{g}^{\delta}(i) = \sum_{a \in \mathcal{A}(i)} [\delta(i)]_a \, g(i, a),$$

where $g(i, a) = \mathrm{E}\,[g_t \mid s_t = i, u_t = a]$ is the expected immediate cost of taking action a from state i. Let P_{δ} denote the $n \times n$ substochastic matrix given by $[P_{\delta}]_{ij} = p_{ij}^{\delta}$

for $i, j \in \{1, 2, \ldots, n\}$. Let $\bar{g}^\delta \in \mathbf{R}^n$ be the expected immediate cost vector whose i^{th} component is $\bar{g}^\delta(i)$, for $i \in \{1, 2, \ldots, n\}$.

For each stationary randomized policy $\delta \in \Lambda$, the cost to go function for the SSP problem is

$$\tilde{J}^\delta = (I - P_\delta)^{-1} \bar{g}^\delta = \sum_{l=0}^{\infty} P_\delta^l \bar{g}^\delta$$

where P_δ^l is P_δ multiplied with itself l times, and $P_\delta^0 = I$, the $n \times n$ identity matrix. Let

$$\mathcal{Q} \equiv \{(i, a) \mid i \in \{1, 2, \ldots, n\}, a \in \mathcal{A}(i)\}.$$

We introduce the function $\mathsf{h} : (i, a, V) \mapsto \mathbf{R}$ as follows :

$$\mathsf{h}(i, a, V) = g(i, a) + \sum_{j=1}^{n} p_{ij}(a) V(j)$$

for $i \in \{1, 2, \ldots, n\}$, $a \in \mathcal{A}(i)$, $V \in \mathbf{R}^n$. For each $J \in \mathbf{R}^n$, let the operator $\tilde{T}_\delta : \mathbf{R}^n \to \mathbf{R}^n$ be defined by

$$\begin{aligned} \left(\tilde{T}_\delta J\right)(i) &= \bar{g}^\delta(i) + \sum_{j=1}^{n} p_{ij}^\delta \, J(j) \\ &= \sum_{a \in \mathcal{A}(i)} [\delta(i)]_a \, \mathsf{h}(i, a, J). \end{aligned}$$

Let $\| \cdot \|$ denote the sup-norm or ℓ_∞ norm defined by

$$\| J \| = \max_{i \in \{1, 2, \ldots, n\}} |J(i)|$$

for $J \in \mathbf{R}^n$. Then $\tilde{T}_\delta$ is a monotone operator which is also a non-expansion with respect to the sup-norm (see [12] and Appendix C). $\tilde{T}_\delta$ is also an n-stage contraction mapping under the sup-norm (see Appendix C). In fact, $\tilde{J}^\delta$ is the unique fixed point

of the operator $\tilde{T}_\delta$. Thus for $i \in \{1, 2, \ldots, n\}$,

$$\tilde{J}^\delta(i) = \sum_{a \in \mathcal{A}(i)} [\delta(i)]_a \underbrace{\left(g(i,a) + \sum_{j=1}^{n} p_{ij}(a) \tilde{J}^\delta(j) \right)}_{Q^\delta(i,a)}.$$

It is easy to see from the definition that $Q^\delta(i,a)$ is the expected total cost of taking action a from state i at time $t = 0$, and from then on following the policy δ. $Q^\delta(i,a) = \mathsf{h}(i, a, \tilde{J}^\delta)$ for $(i,a) \in \mathcal{Q}$. Let $\tilde{J}^* \in \mathbf{R}^n$ denote the optimal (minimal) cost to go function for the SSP problem, also $Q^*(i,a) = \mathsf{h}(i, a, \tilde{J}^*)$ denotes the optimal Q-values.

Note that for any $J \in \mathbf{R}^n$ and (proper) $\delta \in \Lambda$, $\lim_{l \to \infty} \tilde{T}_\delta^l J = \tilde{J}^\delta$. Here $\tilde{T}_\delta^l$ is the composition of the operator $\tilde{T}_\delta$ with itself l times; i.e. $\tilde{T}_\delta^{(l+1)} J = \tilde{T}_\delta(\tilde{T}_\delta^l J)$ for $J \in \mathbf{R}^n$ and $l \geq 0$. $\tilde{T}_\delta^0$ is the identity operator, i.e. $\tilde{T}_\delta^0 J = J$.

Note that the dynamic programming operator $\tilde{T} : \mathbf{R}^n \to \mathbf{R}^n$ is given by

$$(\tilde{T}J)(i) = \min_{a \in \mathcal{A}(i)} \mathsf{h}(i, a, J) \qquad \text{for } i \in \{1, 2, \ldots, n\}$$

for $J \in \mathbf{R}^n$. $\tilde{T}$ is a monotone operator which is a non-expansion under the sup-norm (see Appendix C). $\tilde{T}$ is an n-stage contraction mapping under the sup-norm and $\tilde{J}^*$, the optimal cost to go function for the SSP problem, is its unique fixed point. Note that for any $J \in \mathbf{R}^n$, $\lim_{l \to \infty} \tilde{T}^l J = \tilde{J}^*$. Here $\tilde{T}^l$ is the composition of the operator $\tilde{T}$ with itself l times; i.e. $\tilde{T}^{(l+1)} J = \tilde{T}(\tilde{T}^l J)$ for $J \in \mathbf{R}^n$ and $l \geq 0$, also $\tilde{T}^0$ is the identity operator, i.e. $\tilde{T}^0 J = J$.

Let $\tilde{\delta}$ be another stationary randomized policy (proper) such that

$$\sum_{a \in \mathcal{A}(i)} [\tilde{\delta}(i)]_a \, Q^\delta(i,a) \leq \tilde{J}^\delta(i) \qquad \text{for } i \in \{1, 2, \ldots, n\}.$$

Then it follows from the monotonicity property (see Appendix C) of the operator $\tilde{T}_{\tilde{\delta}}$ that $\tilde{J}^{\tilde{\delta}} \leq \tilde{J}^{\delta}$. Let the scalars $\alpha_i > 0$, for $i \in \{1, 2, \ldots, n\}$. It follows that any local minimum of $\mathsf{s}(\delta) \equiv \sum_{i=1}^{n} \alpha_i \tilde{J}^{\delta}(i)$ is also a global minimum of $\mathsf{s}(\delta)$ in the domain Λ.

Let

$$\Lambda_{\bar{\epsilon}} \equiv \{\delta \in \Lambda \mid [\delta(i)]_a \geq \bar{\epsilon}(i),\ i \in \{1, 2, \ldots, n\},\ a \in \mathcal{A}(i)\}.$$

where $\bar{\epsilon} \in \mathbf{R}^n$, with $\bar{\epsilon}(i) \geq 0$ for $i \in \{1, 2, \ldots, n\}$. Here $\bar{\epsilon}(i)$ denotes the i^{th} component of $\bar{\epsilon}$. Let $\tilde{\epsilon} \in \mathbf{R}^n$ be the vector with

$$\tilde{\epsilon}(i) = \frac{1}{|\mathcal{A}(i)|} \qquad \text{for } i \in \{1, 2, \ldots, n\}.$$

Then $\underline{0} \leq \bar{\epsilon} \leq \tilde{\epsilon}$ implies that $\Lambda_{\bar{\epsilon}}$ is non-empty, here $\underline{0} \in \mathbf{R}^n$ is the vector with all components equal to zero and the inequality is componentwise. Also, $\underline{0} \leq \bar{\epsilon} \leq \hat{\epsilon} \leq \tilde{\epsilon}$ imply that $\Lambda_{\hat{\epsilon}} \subset \Lambda_{\bar{\epsilon}}$.

For each positive integer k and scalar ϵ, where $0 \leq \epsilon \leq \frac{1}{k}$, define

$$\Delta_k^{\epsilon} \equiv \{(p_1, p_2, \ldots, p_k) \mid p_l \geq \epsilon, \sum_{l=1}^{k} p_l = 1\}.$$

We define the k extremal points of Δ_k^{ϵ} (when $0 \leq \epsilon < \frac{1}{k}$) as follows: the i^{th} one is defined as the probability vector $(p_1, p_2, \ldots, p_k)$ with

$$\begin{aligned} p_i &= (1 - (k-1)\epsilon) \\ p_j &= \epsilon, \qquad \text{when } j \neq i \end{aligned}$$

Note that when $k > 1$, $p_i > p_j,\ j \neq i$. Also $\Lambda_{\underline{0}} = \Lambda$. A $\delta \in \Lambda_{\bar{\epsilon}}$ with $\underline{0} \leq \bar{\epsilon} < \tilde{\epsilon}$ is called an extremal policy of $\Lambda_{\bar{\epsilon}}$ if $\delta(i)$ is an extremal point of $\Delta_{|\mathcal{A}(i)|}^{\bar{\epsilon}(i)}$ for each $i \in \{1, 2, \ldots, n\}$. The strict inequality holds componentwise.

Observe that the extremal policies of $\Lambda_{\underline{0}}$ are precisely the stationary deterministic policies. Let Υ denote the set of stationary deterministic policies (or control functions to be precise). We use the notation μ exclusively to denote stationary deterministic policies; $\mu(i) \in \mathcal{A}(i)$ for $i \in \{1, 2, \ldots, n\}$. As mentioned earlier it is implicitly assumed that the action taken from state 0 for any stationary deterministic policy is the unique action in $\mathcal{A}(0)$, namely 1. Note that there is a natural one to one correspondence between the elements of Υ and the extremal policies of $\Lambda_{\bar{\epsilon}}$ when $\underline{0} \leq \bar{\epsilon} < \tilde{\epsilon}$. An extremal policy δ of $\Lambda_{\bar{\epsilon}}$ corresponding to a stationary deterministic policy $\mu \in \Upsilon$ has the property that

$$[\delta(i)]_{\mu(i)} > [\delta(i)]_{a'} \qquad \text{if } a' \in \mathcal{A}(i),\ a' \neq \mu(i)$$

for $i \in \{1, 2, \ldots, n\}$. Without loss of generality we will use $\mu \in \Upsilon$ to denote either the extremal policies of $\Lambda_{\underline{0}}$ or the corresponding control law, mapping the states $\{1, 2, \ldots, n\}$ to the corresponding action in each state on which all the probability mass is concentrated. It will be clear from the context whether $\mu(i),\ i \in \{1, 2, \ldots, n\}$, denotes an extremal point of $\Delta^0_{|\mathcal{A}(i)|}$ or the corresponding control action in $\mathcal{A}(i)$.

Just in this appendix, for any positive integer k and any $w \in \mathbf{R}^k$, let $\| w \|_1$ denote the ℓ_1 norm defined by $\sum_{l=1}^k |w_l|$. Here w_l is the l^{th} component of w. We define a metric d on the set Λ as follows. For any $\delta, \tilde{\delta} \in \Lambda$, define

$$\mathsf{d}(\delta, \tilde{\delta}) = \max_{i \in \{1,2,\ldots,n\}} \| \delta(i) - \tilde{\delta}(i) \|_1 \ .$$

It is easy to see that this is a metric and that Λ is a compact space under this metric.

Define

$$\text{Interior}(\Lambda) \equiv \bigcup_{\bar{\epsilon}:\underline{0}<\bar{\epsilon}\leq\tilde{\epsilon}} \Lambda_{\bar{\epsilon}}.$$

Note that δ is an element of Interior(Λ) if and only if δ assigns positive probability to each possible action from each state $i \in \{1, 2, \ldots, n\}$. Such a δ is called a stationary fully randomized policy. Also note that the components of P_δ and $\bar{g}^\delta$ are continuous functions on the space Λ. Note that the cost to go function for policy $\delta \in \Lambda$, given by $\tilde{J}^\delta = (I - P_\delta)^{-1} \bar{g}^\delta$, is a continuous function on the space Λ. In fact, the compactness of Λ implies that $\tilde{J}^\delta$ is uniformly continuous on Λ. In particular, given a scalar $\epsilon > 0$, there exists a scalar $\varsigma > 0$ (dependent on ϵ) such that $\| \tilde{J}^\mu - \tilde{J}^\delta \| < \epsilon$, for each $\mu \in \Upsilon$ and $\delta \in \Lambda$ with $\mathsf{d}(\mu, \delta) < \varsigma$.

A policy $\mu \in \Upsilon$ is said to be a *greedy policy* for $V \in \mathbf{R}^n$ if

$$\mu(i) = \arg \min_{a \in \mathcal{A}(i)} \mathsf{h}(i, a, V) \qquad \text{for } i \in \{1, 2, \ldots, n\}.$$

Note that for each $i \in \{1, 2, \ldots, n\}$ and $a \in \mathcal{A}(i)$, the function $\mathsf{h}(i, a, \cdot)$ is an affine function on the space $\mathbf{R}^n$. Note that for any $\mu \in \Upsilon$, the operator $\tilde{T}_\mu$ is such that $(\tilde{T}_\mu V)(i) = \mathsf{h}(i, \mu(i), V)$ for $V \in \mathbf{R}^n$ and $i \in \{1, 2, \ldots, n\}$. We define for each $\mu \in \Upsilon$, the "greedy region" for μ as

$$\tilde{\mathcal{R}}_\mu \equiv \{V \in \mathbf{R}^n \mid \mu \text{ is greedy for } V\}$$

It is easy to see that $\tilde{\mathcal{R}}_\mu$ is a polyhedron. Also note that $\tilde{\mathcal{R}}_\mu$ may be empty for some μ and that $\mathbf{R}^n = \bigcup_{\mu \in \Upsilon} \tilde{\mathcal{R}}_\mu$. Since a policy $\mu \in \Upsilon$ is optimal if and only if $\tilde{T}_\mu \tilde{J}^\mu = \tilde{T} \tilde{J}^\mu$ (see [12]), a policy $\mu \in \Upsilon$ is optimal if and only if $\tilde{J}^\mu \in \tilde{\mathcal{R}}_\mu$. In fact, such an optimal policy $\mu^* \in \Upsilon$ exists [11, 12].

D.2 Approximate Policy Iteration

For $V \in \mathbf{R}^n$, let $\tilde{\mathsf{h}}(i,a,V)$ denote an approximation to $\mathsf{h}(i,a,V)$ for each $i \in \{1,2,\ldots,n\}$ and $a \in \mathcal{A}(i)$.

Lemma D.1 *Let $\tilde{V}$ be any fixed vector in $\mathbf{R}^n$. Then there exist scalars $\epsilon > 0$, $\varsigma > 0$ dependent on $\tilde{V}$, such that if V is any vector in $\mathbf{R}^n$ with $\| V - \tilde{V} \| < \epsilon$ and $\tilde{\mathsf{h}}$ is such that $|\mathsf{h}(i,a,V) - \tilde{\mathsf{h}}(i,a,V)| < \varsigma$ for all $i \in \{1,2,\ldots,n\}$ and $a \in \mathcal{A}(i)$, then the control policy $\tilde{\mu} \in \Upsilon$ obtained by setting $\tilde{\mu}(i) = \arg\min_{a \in \mathcal{A}(i)} \tilde{\mathsf{h}}(i,a,V)$ for each $i \in \{1,2,\ldots,n\}$ is a greedy policy for the vector $\tilde{V}$.*

□

The proof of the above lemma follows from the affine nature of $\mathsf{h}(i,a,\cdot)$ and the finiteness of the number of states and actions. See also Chapter 4.

Corollary D.1 *Consider an SSP problem. Fix a $\mu \in \Upsilon$. There exist scalars $\epsilon > 0$, $\varsigma > 0$, such that if J is any vector in $\mathbf{R}^n$ with $\| J - \tilde{J}^{\mu} \| < \epsilon$ and $|\tilde{\mathsf{h}}(i,a,J) - \mathsf{h}(i,a,J)| < \varsigma$ for $i \in \{1,2,\ldots,n\}$, $a \in \mathcal{A}(i)$, then the control policy $\tilde{\mu} \in \Upsilon$ obtained by setting $\tilde{\mu}(i) = \arg\min_{a \in \mathcal{A}(i)} \tilde{\mathsf{h}}(i,a,J)$ for each $i \in \{1,2,\ldots,n\}$ is a greedy policy for the vector $\tilde{J}^{\mu}$. In fact, the ϵ and ς are applicable uniformly to all $\mu \in \Upsilon$.*

□

The proof of the above corollary follows from the fact that the cardinality of Υ, namely $|\Upsilon|$, is finite and from Lemma D.1.

Let

$$\tilde{\rho}_n \equiv \max_{i\in\{1,2,\ldots,n\}} \max_{\mu\in\Upsilon} \Pr^{\mu}\left[s_n \neq 0 \mid s_0 = i\right].$$

Here $\Pr^{\mu}\left[\cdot \mid i\right]$ denotes the probability measure induced by stationary policy μ, given the starting state $s_0 = i$. See Appendix C for more on notations. See Appendix C for the definition of $\tilde{\rho}_k$ for non-negative integer k. Note that $0 \leq \tilde{\rho}_n < 1$ since we assume all stationary deterministic policies to be proper.

Assume that a sequence of stationary deterministic policies $\{\mu_k\}$ and a corresponding sequence of approximate cost to go functions $\{J_k\}$, where $J_k \in \mathbf{R}^n$, satisfy

$$\max_{i\in\{1,2,\ldots,n\}} |J_k(i) - \tilde{J}^{\mu_k}(i)| \;\leq\; \varepsilon, \qquad k = 0, 1, 2, \ldots$$

and

$$\max_{i\in\{1,2,\ldots,n\}} |(\tilde{T}_{\mu_{k+1}} J_k)(i) - (\tilde{T} J_k)(i)| \;\leq\; \epsilon, \qquad k = 0, 1, 2, \ldots$$

Then (see [12] and Appendix C)

$$\limsup_{k\to\infty} \max_{i\in\{1,2,\ldots,n\}} \left(\tilde{J}^{\mu_k}(i) - \tilde{J}^{*}(i)\right) \leq \frac{n(1-\tilde{\rho}_n+n)(\epsilon+2\varepsilon)}{(1-\tilde{\rho}_n)^2}.$$

With slight abuse of notation we define $Q^J(i,a) \equiv \mathsf{h}(i,a,J)$ for $J \in \mathbf{R}^n$, $i \in \{1,2,\ldots,n\}$ and $a \in \mathcal{A}(i)$. Hence for $\delta \in \Lambda$, $Q^{\delta}(i,a) = Q^{\tilde{J}^{\delta}}(i,a)$, $(i,a) \in \mathcal{Q}$.

Consider the following algorithm. Pick some $\mu_0 \in \Upsilon$. The sequence $\{\mu_k\}$ of stationary deterministic policies is generated as follows. Let $\{J_k\}$, with $J_k \in \mathbf{R}^n$, be a sequence of vectors generated in such a manner that

$$\| J_k - \tilde{J}^{\mu_k} \| \leq \varepsilon_k, \qquad k = 0, 1, 2, \ldots$$

Let $Q^{J_k}(i,a) \equiv \mathsf{h}(i,a,J_k)$ for $(i,a) \in \mathcal{Q}$. Let $\tilde{Q}_k(i,a),\ (i,a) \in \mathcal{Q}$ be such that

$$|\tilde{Q}_k(i,a) - Q^{J_k}(i,a)| \leq \varsigma_k, \qquad (i,a) \in \mathcal{Q}.$$

We set

$$\mu_{k+1}(i) = \arg\min_{a \in \mathcal{A}(i)} \tilde{Q}_k(i,a) \qquad \text{for } i \in \{1,2,\ldots,n\}.$$

Note that

$$\max_{i \in \{1,2,\ldots,n\}} |(\tilde{T}_{\mu_{k+1}} J_k)(i) - (\tilde{T} J_k)(i)| \leq 2\varsigma_k.$$

Hence we have the following theorem.

Theorem D.1 *Suppose* $\varepsilon = \limsup_{k\to\infty} \varepsilon_k$ *and* $\varsigma = \limsup_{k\to\infty} \varsigma_k$. *Then*

$$\limsup_{k\to\infty} \| \tilde{J}^{\mu_k} - \tilde{J}^* \| \ \leq \ \frac{2n(1-\tilde{\rho}_n + n)(\varsigma + \varepsilon)}{(1-\tilde{\rho}_n)^2}. \tag{D.1}$$

□

Here $\| \cdot \|$ is the sup-norm. We may choose J_k to be equal to $\tilde{J}^{\delta_k}$ when we approximate $\tilde{J}^{\mu_k}$ with $\tilde{J}^{\delta_k}$ for some δ_k close to μ_k (under the metric d defined earlier). Note that if ς and ε are sufficiently small, then $\tilde{J}^{\mu_k} = \tilde{J}^*$ for all large k. This happens when the right hand side of the inequality D.1, is less than

$$\max_{\tilde{\mu} \in \Upsilon,\ \tilde{\mu} \text{ not optimal}} \| \tilde{J}^{\tilde{\mu}} - \tilde{J}^* \|$$

Observe that by Corollary D.1, there exists $\varepsilon > 0$ and $\varsigma > 0$ such that if $\varepsilon_k < \varepsilon$ and $\varsigma_k < \varsigma$ for all $k = 0, 1, 2, \ldots$, then the μ_k's obtained are same as the ones obtained while doing policy iteration (see [12] and Chapter 1), and converges to the optimal policy in a finite ($\leq |\Upsilon|$) number of steps.

D.3 Off-Line Temporal Difference Method For A Proper Policy With Lookup Table Representation

In this section we consider off-line temporal difference (TD) methods [16, Chapter 5] for proper policies. We fix a proper stationary randomized policy $\delta \in \Lambda$, in general and try to compute the cost to go function $\tilde{J}^{\delta} = (I - P_{\delta})^{-1}\bar{g}^{\delta} = \sum_{k=0}^{\infty} P_{\delta}^{k}\, \bar{g}^{\delta}$ using simulation. We assume that the immediate costs have finite variance as mentioned earlier. In this section, from henceforth we drop the subscript and superscript δ associated with the policy.

We use a discrete variable t to index the simulated trajectories that are generated by the algorithm. Let $\mathcal{F}_t$ represent the history of the algorithm upto the point at which the simulation of the t^{th} trajectory is to commence and let $J_t \in \mathbf{R}^n$ be the estimate of the cost to go vector available at that time.

Based on $\mathcal{F}_t$, we choose the initial state i_0^t of the t^{th} trajectory and the step sizes $\gamma_t(i),\ i = 1, \ldots, n$, that will be used for updating '$J(i)$'. We generate a trajectory of states $i_0^t, i_1^t, \ldots, i_{N_t}^t$ under the proper stationary policy (δ), where N_t is the first time that the trajectory reaches state 0. In general N_t may be any stopping time [16] (which may be taken without loss of generality to be less than or equal to the first time that the trajectory reaches state 0). Note that $E\left[N_t \geq k \mid \mathcal{F}_t\right] \leq K\rho^k$ where K and ρ are non-negative scalars, with $0 \leq \rho < 1$. We then update J_t by letting

$$J_{t+1}(i) = J_t(i) + \gamma_t(i) \sum_{m=0}^{N_t - 1} z_m^t(i)\, d_{m,t} \tag{D.2}$$

where the temporal differences $d_{m,t}$ are defined by

$$d_{m,t} = g_{m,t} + J_t(i^t_{m+1}) - J_t(i^t_m)$$

Here $g_{m,t}$ is the immediate cost incurred at the m^{th} instant (or stage) in the t^{th} trajectory (when the action $a^t_m \in \mathcal{A}(i^t_m)$ is taken from the state i^t_m under policy δ). With slight abuse of notation, $J_t(0)$ is assumed to be zero. The initial values $J_1(i),\ \forall i \in \{1, 2, \ldots, n\}$ may be arbitrary. $z^t_m(i)$ are the eligibility coefficients which are assumed to have the following properties.

Assumption D.1 For all m and t and $i \in \{1, 2, \ldots, n\}$ we have

a. $z^t_m(i) \geq 0$.

b. $z^t_{-1}(i) = 0$.

c. $z^t_m(i) \leq z^t_{m-1}(i), \qquad \text{if } \ i^t_m \neq i.$

d. $z^t_m(i) \leq z^t_{m-1}(i) + 1, \qquad \text{if } \ i^t_m = i.$

e. $z^t_m(i)$ is completely determined by $\mathcal{F}_t$ and $i^t_0, \ldots, i^t_m$.

□

Section D.3.1 discusses the choice of the eligibility coefficients. We allow the possibility that no update of $J(i)$ is carried out even if a trajectory visits state i. However for $J(i)$ to converge to the correct value, there should be enough trajectories that lead to a non-trivial update of $J(i)$. For this reason an additional assumption is needed. To this effect we define

$$q_t(i) = \Pr\left[\text{there exists } m \text{ such that } z^t_m(i) > 0 \mid \mathcal{F}_t\right]$$

Here 'Pr' denotes probability. Note that $q_t(i)$ is a function of the past history. We define

$$\mathcal{T}^i = \{t \mid q_t(i) > 0\}$$

which corresponds to the set of trajectories that have a chance of leading to a non-zero update of $J(i)$. Observe that whether t belongs to $\mathcal{T}^i$ or not is only a function of the past history $\mathcal{F}_t$. We now introduce the following assumption.

Assumption D.2

a. For any fixed i and t, $z_m^t(i)$ must be equal to 1, the first time that it becomes positive.

b. There exists a deterministic constant $\kappa > 0$ such that $q_t(i) \geq \kappa$ for all $t \in \mathcal{T}^i$ and all i.

c. $\gamma_t(i) \geq 0$ for all $t \in \mathcal{T}^i$ and $\gamma_t(i) = 0$ for $t \notin \mathcal{T}^i$.

d. $\sum_{t \in \mathcal{T}^i} \gamma_t(i) = \infty$, for all i.

e. $\sum_{t \in \mathcal{T}^i} \gamma_t^2(i) < \infty$, for all i.

□

Section D.5 discusses some aspects of step size selection. Actually, since no update of the i^{th} component $J(i)$ happens when $t \notin \mathcal{T}^i$, whether $\gamma_t(i)$ is zero or not is irrelevant when $t \notin \mathcal{T}^i$.

Proposition D.1 *Consider the off-line temporal difference algorithm, as described by equation D.2 and let Assumption D.1 and Assumption D.2 hold. Assume the*

policy (δ) under consideration is proper. Then $J_t(i)$ converges to $\tilde{J}^{\delta}(i)$ for all $i \in \{1, 2, \ldots, n\}$ almost surely.

□

Though [16] gives the proof of the above proposition for the case when the immediate cost depends only on the current state and subsequent state, the result can easily be shown to hold for the case where the immediate costs are random with finite variance (or finite second moments). We omit the details.

D.3.1 Choice Of Eligibility Coefficients

Suppose that a sample trajectory $i_0, i_1, \ldots$ has been generated. We suppress the index of the t^{th} trajectory, t for convenience. Let $d_m = g_m + J(i_{m+1}) - J(i_m)$, be the temporal difference at the m^{th} stage of the t^{th} trajectory, g_m being the immediate cost at stage m. Similarly, let $z_m(i)$ be the eligibility coefficient in the update

$$J(i) \leftarrow J(i) + \gamma \sum_{m=0}^{\infty} z_m(i)\, d_m$$

Actually $z_m(i)\, d_m = 0$ for $m \geq N$ where N is the stopping time. γ is the step size.

Let us concentrate on a particular state i, and let $m_1, m_2, \ldots m_M$ be the different times that the trajectory is at state i, with M being the total number of such visits. We also use the convention that $m_{M+1} = \infty$. In TD(λ) [16] a temporal difference d_m may lead to an update of $J(i)$ only if i has already been visited by the time m. For this reason, in all our examples we assume $z_m(i) = 0$ for $m < m_1$.

Let $\lambda \in [0, 1]$. We follow the notation $0^0 = 1$. We have the following TD(λ) methods.

a. If we let

$$z_m(i) = \lambda^{m-m_1}, \qquad \text{if } m \geq m_1$$

we have the first visit TD(λ) method.

b. If we let

$$z_m(i) = \sum_{\{j|m_j \leq m\}} \lambda^{m-m_j}$$

we have the every visit TD(λ) method.

c. Consider the choice

$$z_m(i) = \lambda^{m-m_j}, \qquad \text{if } \ m_j \leq m < m_{j+1} \ \ \forall j$$

This gives the restart variant of TD(λ). Note that for $\lambda = 1$, the restart method coincides with the first visit method, whereas for $\lambda = 0$ it coincides with the every visit method.

d. Let us define the stopping time as a random variable τ such that the event $\{\tau \leq k\}$ is completely determined by the history of our simulation upto and including the point that the state i_k is generated. Intuitively the decision whether or not to stop at state i_k must be made before generating subsequent states in a simulated trajectory. Given a stopping time τ, we let

$$z_m(i) = \lambda^{m-m_1}, \qquad \text{for } m_1 \leq m < \tau,$$

and $z_m(i) = 0$ for $m \geq \tau$.

Notice that Assumption D.1 is satisfied by the choices of TD(λ). For other related variants of TD(λ) see [16].

D.4 On-Line Temporal Difference Method For A Proper Policy With Lookup Table Representation

In this section we consider on-line temporal difference methods [16, Chapter 5] for proper policies. We fix a proper stationary randomized policy $\delta \in \Lambda$, in general and try to compute the cost to go function $\tilde{J}^{\delta} = (I - P_{\delta})^{-1}\bar{g}^{\delta} = \sum_{k=0}^{\infty} P_{\delta}^{k}\,\bar{g}^{\delta}$ using simulation. We assume that the immediate costs have finite variance as mentioned earlier. In this section, from henceforth we drop the subscript and superscript δ associated with the policy.

We again use a discrete variable t to index the simulated trajectories that are generated by the algorithm. Let $\mathcal{F}_t$ represent the history of the algorithm upto the point at which the simulation of the t^{th} trajectory is to commence and let $J_t^0 \in \mathbf{R}^n$ be the estimate of the cost to go vector available at the beginning of the t^{th} trajectory. The initial estimates $J_1^0(i),\ \forall i \in \{1, 2, \ldots, n\}$ may be arbitrary.

Based on $\mathcal{F}_t$, we choose the initial state i_0^t of the t^{th} trajectory and the step sizes $\gamma_t(i),\ i = 1, \ldots, n$, that will be used for updating '$J(i)$'. We generate a trajectory of states $i_0^t, i_1^t, \ldots, i_{N_t}^t$ under the proper stationary policy (δ), where N_t is the first time that the trajectory reaches state 0. In general N_t may be any stopping time [16] (which may be taken without loss of generality to be less than or equal to the first time that the trajectory reaches state 0). Note that $E\left[N_t \geq k \mid \mathcal{F}_t\right] \leq K\rho^k$ where K and ρ are non-negative scalars, with $0 \leq \rho < 1$.

Let $J_{t,m}^0 \in \mathbf{R}^n$ be the vector obtained after simulating m transitions of the t^{th}

trajectory. The update equations are as follows.

$$\left.\begin{aligned} J^0_{t,0}(i) &= J^0_t(i), \qquad \forall i \in \{1,2,\ldots,n\} \\ d^0_{m,t} &= g_{m,t} + J^0_{t,m}(i^t_{m+1}) - J^0_{t,m}(i^t_m) \\ J^0_{t,m+1}(i) &= J^0_{t,m}(i) + \gamma_t(i)\, z^t_m(i)\, d^0_{m,t}, \qquad \forall i \in \{1,2,\ldots,n\} \\ J^0_{t+1}(i) &= J^0_{t,N_t}(i), \qquad \forall i \in \{1,2,\ldots,n\} \end{aligned}\right\} \tag{D.3}$$

Note that $g_{m,t}$ is the immediate cost incurred at stage m of the t^{th} trajectory while taking action a^t_m from state i^t_m under the policy δ. The superscript 0 is used in the above equations to indicate that we are dealing with the on-line algorithm. As mentioned earlier N_t is the length of the t^{th} trajectory. Note that the step sizes $\gamma_t(i)$ are held constant during each trajectory. We then have the following convergence result [16]

Proposition D.2 *Consider the on-line temporal difference algorithm, as described in equations D.3 and let Assumption D.1 and Assumption D.2 hold. Furthermore assume that the eligibility coefficients $z^t_m(i)$ are bounded by a deterministic constant C. Assume that the policy (δ) under consideration is proper. Then $J^0_t(i)$ converges to $\tilde{J}^\delta(i)$ for all $i \in \{1,2,\ldots,n\}$ almost surely.*

□

The assumption that $z^t_m(i)$ is bounded is satisfied whenever we are dealing with the first visit or the restart variant of TD(λ), because $z^t_m(i)$ is bounded above by 1. Also if $\lambda < 1$, it is easily seen that under every visit TD(λ) method we have $z^t_m(i) \leq \frac{1}{(1-\lambda)}$ and our assumption is again satisfied.

Though [16] gives the proof of the above proposition for the case when the immediate cost depends only on the current state and subsequent state, the result can easily be shown to hold for the case where the immediate costs are random with finite variance (or finite second moments). We omit the details.

Notice that it may be inferred from the proof of the above proposition [16, Chapter 5, Section 5.3.6] that

$$\max_{i\in\{1,2,\ldots,n\}} \max_{m\in\{0,1,2,\ldots,N_t\}} |J^0_{t,m}(i) - \tilde{J}^{\delta}(i)| \stackrel{t\to\infty}{\longrightarrow} 0$$

almost surely under the assumptions in Proposition D.2. We however omit the details of the proof of this inference.

D.5 A Remark On Step Size Selection

To ensure Assumptions D.2 (d)-(e), we might need to know whether $t \in \mathcal{T}^i$ or not which may be non trivial. Please refer [16] for details.

An alternative is as follows [16]. Let for each $i \in \{1, 2, \ldots, n\}$, $\{\tilde{\gamma}_k(i)\}$ be a deterministic non-negative sequence such that $\sum_{k=1}^{\infty} \tilde{\gamma}_k(i) = \infty$, $\sum_{k=1}^{\infty} (\tilde{\gamma}_k(i))^2 < \infty$ for all $i \in \{1, 2, \ldots, n\}$. Now choose $\gamma_t(i) = \tilde{\gamma}_{k+1}(i)$ if there have been exactly k past trajectories during which '$J(i)$' was updated; that is there have been exactly k past trajectories during which the eligibility coefficient '$z(i)$' became positive. This step size rule does satisfy Assumptions D.2 (d)-(e). We may prove this fact using an argument exactly along the lines of [16, page 218]. We omit the details.

We need to assume that the eligibility coefficient '$z(i)$' becomes positive for an infinite number of trajectories, i.e. '$J(i)$' is updated an infinite number of times.

This assumption is natural and is clearly necessary in order to prove convergence.

D.6 Convergence For Discounted Cost Problems

See Chapter 1 and Chapter 4 for a discussion on discounted cost problems. For a fixed stationary randomized policy $\delta \in \Lambda$ we are interested in obtaining the infinite horizon discounted cost to go vector $J^\delta \in \mathbf{R}^n$ given by

$$J^\delta = (I - \beta P_\delta)^{-1} \bar{g}^\delta = \sum_{k=0}^{\infty} (\beta^k P_\delta^k) \bar{g}^\delta$$

using temporal difference methods, where $\beta \in [0, 1)$ is the discount factor. Here the state space is $\{1, 2, \ldots, n\}$. For the state i, $\mathcal{A}(i) = \{1, 2, \ldots, |\mathcal{A}(i)|\}$ denotes the control constraint set for state i. Here $|\mathcal{A}(i)|$ denotes the cardinality of the set $\mathcal{A}(i)$. P_δ is the $n \times n$ transition probability matrix (a stochastic matrix) under the stationary randomized policy δ. i.e.

$$[P_\delta]_{ij} = \sum_{a \in \mathcal{A}(i)} [\delta(i)]_a \, p_{ij}(a)$$

for $i, j \in \{1, 2, \ldots, n\}$. $[\delta(i)]_a$ is the probability of taking action a from state i under policy δ. $p_{ij}(a)$ is the probability that the next state is j given current state is i and action taken is $a \in \mathcal{A}(i)$. Also

$$\bar{g}^\delta(i) = \sum_{a \in \mathcal{A}(i)} [\delta(i)]_a \, g(i, a)$$

is the expected immediate cost from state i under policy δ. Here $g(i, a)$ is the expected immediate cost of taking action a from state i. We assume this immediate costs to have finite variance.

The first method for temporal difference scheme is to consider the equivalent Stochastic Shortest Path Problem (see Appendix C and [16]) and use the results of TD learning for the SSP as discussed in the previous sections.

In a second alternative [16, Section 5.3.7] we only simulate trajectories for a finite number N_t of time steps, which is tantamount to setting the eligibility coefficients '$z_m^t(i)$' to zero for $m \geq N_t$. In general we may take N_t to be a stopping time.

The main differences that arise in the discounted cost case are as follows. First the discount factor β enters in the definition

$$d_m = g_m + \beta J(i_{m+1}) - J(i_m)$$

of the temporal difference. In the above definition of temporal difference we have suppressed the index of the t^{th} trajectory t, for convenience. To be more precise

$$d_{m,t} = g_{m,t} + \beta\, J_t(i_{m+1}^t) - J_t(i_m^t)$$

for the off-line scheme and

$$d_{m,t}^0 = g_{m,t} + \beta\, J_{t,m}^0(i_{m+1}^t) - J_{t,m}^0(i_m^t)$$

for the on-line scheme. $g_{m,t}$ is the immediate cost incurred at stage m of the t^{th} trajectory when action $a_m^t \in \mathcal{A}(i_m^t)$ is taken from state i_m^t under policy δ.

A second difference is that we replace Assumption D.1 (c) with

$$z_m^t(i) \leq \beta\, z_{m-1}^t(i), \qquad \text{if } \; i_m^t \neq i$$

Note that for the TD(λ) scheme considered in Subsection D.3.1, if we replace λ by $\beta\lambda$ in the definition of the eligibility coefficients (for the discounted cost case), this

assumption is satisfied by them. Furthermore we impose the condition that

$$\Pr\left(N_t \geq k \mid \mathcal{F}_t\right) \leq K\rho^k \qquad \forall k \geq 0,\ t \geq 1$$

where K and ρ are non-negative constants with $\rho < 1$.

Then it may be shown [16, Section 5.3.7] that results similar to Proposition D.1 and Proposition D.2 corresponding to off-line and on-line schemes exists for the discounted cost case.

D.7 TD For Learning

Here we are interested in learning the optimal cost to go function and optimal stationary policy for the "all stationary deterministic policies proper" SSP problem. Neither the transition probabilities nor the distribution of the immmediate costs are known. We use TD schemes for evaluating the cost to go function and estimate Q-values using small step stochastic approximation, employing stationary fully randomized policies which are 'near' to stationary deterministic policies, for exploration.

Assumption D.3

a. *For each state action pair $(i, a) \in \mathcal{Q}$, let the pre-determined scalar non-negative step sizes $\gamma_k(i, a)$ be such that*

$$\sum_{k=1}^{\infty} \gamma_k(i,a) = \infty; \qquad \sum_{k=1}^{\infty} \left(\gamma_k(i,a)\right)^2 < \infty$$

b. *The immmediate costs have finite second moments, i.e.*

$$E\left[g_t^2 \mid s_t = i, u_t = a\right] < \infty, \qquad \forall (i,a) \in \mathcal{Q}$$

Here g_t is the immediate cost (random) of taking action $u_t = a$ from state $s_t = i$.

□

Fix a policy, $\delta \in \text{Interior}(\Lambda)$. We would like to estimate $\tilde{J}^\delta \in \mathbf{R}^n$, the cost to go function for policy δ, for the SSP problem. We would also like to estimate the Q values, namely

$$Q^\delta(i,a) = Q^{\tilde{J}^\delta}(i,a) = g(i,a) + \sum_{j=1}^{n} p_{ij}(a)\, \tilde{J}^\delta(j), \quad \forall (i,a) \in \mathcal{Q}$$

We use the off-line temporal difference scheme to estimate $\tilde{J}^\delta$ in the following algorithm.

Algorithm D.1

Input : Stationary Randomized Policy $\delta \in$ Interior(Λ).

Output : Estimate $\tilde{J}(i)$ of $\tilde{J}^\delta(i)$ for $i \in \{1, 2, \ldots, n\}$ and estimates $\tilde{Q}(i,a)$ of $Q^\delta(i,a) = g(i,a) + \sum_{j=1}^{n} p_{ij}(a)\, \tilde{J}^\delta(j)$, for $(i,a) \in \mathcal{Q}$.

t : index of trajectory; $t \in \{1, 2, \ldots\}$.

m : index of stage within each trajectory; $m \in \{0, 1, 2, \ldots\}$.

$\tilde{n}$: number of trajectories simulated, a positive integer.

N_t : stopping time for t^{th} trajectory.

i_m^t : state at m^{th} stage of t^{th} trajectory.

a_m^t : action taken at m^{th} stage of the t^{th} trajectory, from state i_m^t under policy δ.

$z_m^t(i)$: *eligibility coefficients.*

$\gamma_t(i)$: *step sizes for off-line TD scheme.*

$g_{m,t}$: *immediate cost incurred at stage* m *of* t^{th} *trajectory, when action* a_m^t *is taken from state* i_m^t.

$d_{m,t}$: *temporal difference at stage* m *of* t^{th} *trajectory.*

$Q_{t,m}(i,a)$: *estimate of* $Q^\delta(i,a)$ *at stage* m *of* t^{th} *trajectory.*

$J_t(i)$: *estimate of* $\tilde{J}^\delta(i)$ *at the start of* t^{th} *trajectory.*

$\tau_m^t(i,a)$: *number of times action 'a' has been taken from state 'i', by the time (including stage* m*) stage* m *is reached in the* t^{th} *trajectory.*

1.

$$\begin{aligned} t &= 1 \\ \tau_{-1}^1(i,a) &= 0 \qquad \forall (i,a) \in \mathcal{Q} \end{aligned}$$

The initial values

$$\begin{aligned} Q_{1,0}(i,a) &\qquad \textit{arbitrary} \quad \forall (i,a) \in \mathcal{Q} \\ J_1(i) &\qquad \textit{arbitrary} \quad \forall i \in \{1,2,\ldots,n\} \end{aligned}$$

2. $\mathrm{zd}_t(i) = 0, \quad \forall i \in \{1,2,\ldots,n\}$.

3. For $m = 0$ *to* $N_t - 1$*, do*

$$\tau_m^t(i_m^t, a_m^t) = \tau_{m-1}^t(i_m^t, a_m^t) + 1$$

$$
\begin{aligned}
\tau_m^t(i,a) &= \tau_{m-1}^t(i,a), \quad \forall (i,a) \neq (i_m^t, a_m^t),\ (i,a) \in \mathcal{Q} \\
Q_{t,m+1}(i_m^t, a_m^t) &= \left(1 - \gamma_{\tau_m^t(i_m^t,a_m^t)}(i_m^t, a_m^t)\right) Q_{t,m}(i_m^t, a_m^t) \\
&\quad + \gamma_{\tau_m^t(i_m^t,a_m^t)}(i_m^t, a_m^t) \left(g_{m,t} + J_t(i_{m+1}^t)\right) \qquad \text{(D.4)} \\
Q_{t,m+1}(i,a) &= Q_{t,m}(i,a), \quad \forall (i,a) \neq (i_m^t, a_m^t),\ (i,a) \in \mathcal{Q} \\
d_{m,t} &= g_{m,t} + J_t(i_{m+1}^t) - J_t(i_m^t) \\
\mathrm{zd}_t(i) &= \mathrm{zd}_t(i) + z_m^t(i)\, d_{m,t}, \qquad \forall i \in \{1, 2, \ldots, n\}
\end{aligned}
$$

4.

$$
\begin{aligned}
J_{t+1}(i) &= J_t(i) + \gamma_t(i)\, \mathrm{zd}_t(i) \qquad \forall i \in \{1, 2, \ldots, n\} \\
Q_{t+1,0}(i,a) &= Q_{t,N_t}(i,a) \qquad \forall (i,a) \in \mathcal{Q} \\
\tau_{-1}^{t+1}(i,a) &= \tau_{N_t-1}^t(i,a) \qquad \forall (i,a) \in \mathcal{Q}
\end{aligned}
$$

5. $t = t + 1$.

6. go to step 2, if $t \leq \tilde{n}$*; else go to step 7.*

7. Return

$$
\begin{aligned}
\tilde{Q}(i,a) &= Q_{t,0}(i,a) \qquad \forall (i,a) \in \mathcal{Q} \\
\tilde{J}(i) &= J_t(i) \qquad \forall i \in \{1, 2, \ldots, n\}
\end{aligned}
$$

□

Note that '$J_t(0)$' is defined as zero. We assume Assumption D.1, Assumption D.2 and Assumption D.3 to hold.

All that is required of the non-negative step size parameters $\gamma_k(i,a)$ is that they satisy the standard asssumptions

$$\sum_{k=1}^{\infty} \gamma_k(i,a) = \infty; \quad \sum_{k=1}^{\infty} (\gamma_k(i,a))^2 < \infty$$

for each $(i,a) \in \mathcal{Q}$ almost surely, and may be allowed to depend on the past history, i.e. if the k^{th} time that action 'a' is taken from state 'i', is at the m^{th} stage of the t^{th} trajectory (note that $i_m^t = i,\ a_m^t = a$ in this case), then

$$\gamma_{\tau_m^t(i,a)}(i,a) = \gamma_k(i,a)$$

can depend on the past history until the m^{th} stage of the t^{th} trajectory (after the decision to take action a_m^t is made) but before the action a_m^t is taken at the m^{th} stage of the t^{th} trajectory.

Since Assumption D.1 and Assumption D.2 hold, each state $i \in \{1,2,\ldots,n\}$ is visited infinitely often if Algorithm D.1 is run for an infinite number of trajectories. By Proposition D.1, we have $J_t(i) \stackrel{t\to\infty}{\longrightarrow} \tilde{J}^{\delta}(i)$ for $i \in \{1,2,\ldots,n\}$. Since $\delta \in \text{Interior}(\Lambda)$, each state action pair $(i,a) \in \mathcal{Q}$ is taken infinitely often also. Standard results from stochastic approximation theory [16] can be used to show that under Assumption D.3, we have $Q_{t,0}(i,a) \stackrel{t\to\infty}{\longrightarrow} Q^{\delta}(i,a)$ for $(i,a) \in \mathcal{Q}$.

Now we turn to on-line TD schemes for estimating the cost to go function. Fix a policy $\delta \in \text{Interior}(\Lambda)$. We use on-line TD scheme to estimate $\tilde{J}^{\delta}$, the cost to go function for the SSP problem for policy δ. We would also like to estimate the Q values, namely

$$Q^{\delta}(i,a) = Q^{\tilde{J}^{\delta}}(i,a) = g(i,a) + \sum_{j=1}^{n} p_{ij}(a)\, \tilde{J}^{\delta}(j), \quad \forall (i,a) \in \mathcal{Q}$$

We assume that Assumption D.1, Assumption D.2 and Assumption D.3 hold.

Algorithm D.2

Input : Stationary Randomized Policy $\delta \in$ *Interior*(Λ).

Output : Estimate $\tilde{J}(i)$ *of* $\tilde{J}^{\delta}(i)$ *for* $i \in \{1, 2, \ldots, n\}$ *and estimates* $\tilde{Q}(i,a)$ *of* $Q^{\delta}(i,a) = g(i,a) + \sum_{j=1}^{n} p_{ij}(a)\, \tilde{J}^{\delta}(j)$, *for* $(i,a) \in \mathcal{Q}$.

t *: index of trajectory;* $t \in \{1, 2, \ldots\}$.

m *: index of stage within each trajectory;* $m \in \{0, 1, 2, \ldots\}$.

$\tilde{n}$ *: number of trajectories simulated, a positive integer.*

N_t *: stopping time for* t^{th} *trajectory.*

i_m^t *: state at* m^{th} *stage of* t^{th} *trajectory.*

a_m^t *: action taken at* m^{th} *stage of the* t^{th} *trajectory, from state* i_m^t *under policy* δ.

$z_m^t(i)$ *: eligibility coefficients.*

$\gamma_t(i)$ *: step sizes for on-line TD scheme.*

$g_{m,t}$ *: immediate cost incurred at stage* m *of* t^{th} *trajectory, when action* a_m^t *is taken from state* i_m^t.

$d_{m,t}^0$ *: temporal difference at stage* m *of* t^{th} *trajectory.*

$Q_{t,m}(i,a)$ *: estimate of* $Q^{\delta}(i,a)$ *at stage* m *of* t^{th} *trajectory.*

$J_t^0(i)$ *: estimate of* $\tilde{J}^{\delta}(i)$ *at the start of* t^{th} *trajectory.*

$J^0_{t,m}(i)$: *estimate of* $\tilde{J}^\delta(i)$ *at the* m^{th} *stage of the* t^{th} *trajectory.*

$\tau^t_m(i,a)$: *number of times action 'a' has been taken from state 'i', by the time (including stage* m*) stage* m *is reached in the* t^{th} *trajectory.*

1.

$$\begin{aligned} t &= 1 \\ \tau^1_{-1}(i,a) &= 0 \qquad \forall (i,a) \in \mathcal{Q} \end{aligned}$$

The initial values

$$\begin{aligned} Q_{1,0}(i,a) &\qquad \textit{arbitrary} \quad \forall (i,a) \in \mathcal{Q} \\ J^0_1(i) &\qquad \textit{arbitrary} \quad \forall i \in \{1,2,\ldots,n\} \end{aligned}$$

2. $J^0_{t,0}(i) = J^0_t(i), \quad \forall i \in \{1,2,\ldots,n\}.$

3. For $m = 0$ *to* $N_t - 1$*, do*

$$\begin{aligned} \tau^t_m(i^t_m, a^t_m) &= \tau^t_{m-1}(i^t_m, a^t_m) + 1 \\ \tau^t_m(i,a) &= \tau^t_{m-1}(i,a), \quad \forall (i,a) \neq (i^t_m, a^t_m), \ (i,a) \in \mathcal{Q} \\ d^0_{m,t} &= g_{m,t} + J^0_{t,m}(i^t_{m+1}) - J^0_{t,m}(i^t_m) \\ J^0_{t,m+1}(i) &= J^0_{t,m}(i) + \gamma_t(i)\, z^t_m(i)\, d^0_{m,t}, \quad \forall i \in \{1,2,\ldots,n\} \\ Q_{t,m+1}(i^t_m, a^t_m) &= \left(1 - \gamma_{\tau^t_m(i^t_m,a^t_m)}(i^t_m, a^t_m)\right) Q_{t,m}(i^t_m, a^t_m) \\ &\quad + \gamma_{\tau^t_m(i^t_m,a^t_m)}(i^t_m, a^t_m) \left(g_{m,t} + J^0_{t,m+1}(i^t_{m+1})\right) \qquad \text{(D.5)} \\ Q_{t,m+1}(i,a) &= Q_{t,m}(i,a), \quad \forall (i,a) \neq (i^t_m, a^t_m), \quad (i,a) \in \mathcal{Q} \end{aligned}$$

4.

$$J^0_{t+1}(i) = J^0_{t,N_t}(i) \qquad \forall i \in \{1,2,\ldots,n\}$$

$$
\begin{aligned}
Q_{t+1,0}(i,a) &= Q_{t,N_t}(i,a) \qquad \forall (i,a) \in \mathcal{Q} \\
\tau_{-1}^{t+1}(i,a) &= \tau_{N_t-1}^{t}(i,a) \qquad \forall (i,a) \in \mathcal{Q}
\end{aligned}
$$

5. $t = t + 1$.

6. go to step 2, if $t \leq \tilde{n}$*; else go to step 7.*

7. Return

$$
\begin{aligned}
\tilde{Q}(i,a) &= Q_{t,0}(i,a) \qquad \forall (i,a) \in \mathcal{Q} \\
\tilde{J}(i) &= J_t^0(i) \qquad \forall i \in \{1,2,\ldots,n\}
\end{aligned}
$$

□

The superscript '0' is used to indicate that we are dealing with on-line TD algorithm. The comment following the Algorithm D.1 on the nature of $\gamma_k(i,a)$ is valid in the Algorithm D.2 also. Note that in equation D.5, we could have used $J_t^0(i_{m+1}^t)$ or $J_{t,m}^0(i_{m+1}^t)$ instead of $J_{t,m+1}^0(i_{m+1}^t)$. Note also that '$J_t^0(0)$' and '$J_{t,m}^0(0)$' are defined as zero.

Because of Assumption D.1 and Assumption D.2, each state $i \in \{1,2,\ldots,n\}$ is visited infinitely often if Algorithm D.2 is run for an infinite number of trajectories. By Proposition D.2 we have that $J_t^0(i) \overset{t\to\infty}{\longrightarrow} \tilde{J}^\delta(i)$ for $i \in \{1,2,\ldots,n\}$. Since $\delta \in$ Interior(Λ), each state action pair $(i,a) \in \mathcal{Q}$ is taken infinitely often also. Standard results from stochastic approximation theory [16] can be used to show that under Assumption D.3 we have $Q_{t,0}(i,a) \overset{t\to\infty}{\longrightarrow} Q^\delta(i,a)$ for $(i,a) \in \mathcal{Q}$.

Consider the following algorithm.

Algorithm D.3 *Let $\bar{\epsilon}_k$ be a sequence of positive vectors in $\mathbf{R}^n$ such that $\underline{0} < \bar{\epsilon}_k < \tilde{\epsilon}$, where the inequality is componentwise. Here $\tilde{\epsilon}(i) = \frac{1}{|\mathcal{A}(i)|}$, for $i \in \{1, 2, \ldots, n\}$.*

1. *Set $k = 0$.*

2. *Select an arbitrary stationary randomized policy $\mu_0 \in \Upsilon$.*

3. *Choose the stationary randomized extremal policy $\delta_k \in \Lambda_{\bar{\epsilon}_k}$ associated with μ_k and run Algorithm D.1 (or alternatively Algorithm D.2) for large random number $\tilde{n}_k$ of trajectories, till the cost to go vector "nearly" converges to $\tilde{J}^{\delta_k}$ and the Q values "nearly" converge to $Q^{\delta_k}(i,a) = g(i,a) + \sum_{j=1}^{n} p_{ij}(a)\, \tilde{J}^{\delta_k}$, for $(i,a) \in \mathcal{Q}$.*

   Let $\tilde{J}_k \in \mathbf{R}^n$ be the estimate of the cost to go vector, and $\tilde{Q}_k(i,a),\ \forall (i,a) \in \mathcal{Q}$, be the estimates of the Q values obtained at the end of Algorithm D.1 (or alternatively Algorithm D.2)

   Let

$$\varsigma_k = \max_{(i,a)\in\mathcal{Q}} |Q^{\delta_k}(i,a) - \tilde{Q}_k(i,a)|$$

4. *Set $k = k + 1$, and update the policy μ_k, where*

$$\mu_k(i) = \arg\min_{a\in\mathcal{A}(i)} \tilde{Q}_{k-1}(i,a), \qquad \forall i \in \{1, 2, \ldots, n\}$$

5. *Go to step 3.*

□

Note that we stick with either Algorithm D.1 or Algorithm D.2 in step 3, throughout the execution of Algorithm D.3.

The number of trajectories simulated namely $\tilde{n}_k$, inside the invocation of the Algorithm D.1 (or alternatively Algorithm D.2) in step 3 of Algorithm D.3 may be decided inside the respective algorithm (Algorithm D.1 or alternatively Algorithm D.2) for sufficiently "close" convergence of the cost to go estimate and Q value estimate.

Though the initial estimates of the cost to go vector and Q values in Algorithm D.1 (or alternatively Algorithm D.2) may be arbitrary when called in step 3 (of Algorithm D.3), we may set it to the final estimates in the previous iteration of Algorithm D.3 (i.e. $\tilde{J}_{k-1}$ and $\tilde{Q}_{k-1}(\cdot,\cdot)$).

We have the following theorem.

Theorem D.2 *Consider Algorithm D.3 on an SSP problem where all stationary deterministic policies are proper. Assume that Assumption D.1, Assumption D.2 and Assumption D.3 hold.*

1. *Given any scalar $\epsilon > 0$, there exists an $\bar{\epsilon} \in \mathbf{R}^n$ with $\underline{0} < \bar{\epsilon} < \tilde{\epsilon}$ and a number $\varsigma > 0$, such that if*

$$\limsup_{k\to\infty} \bar{\epsilon}_k(i) \;<\; \bar{\epsilon}(i), \qquad \forall i \in \{1,2,\ldots,n\}$$

and

$$\limsup_{k\to\infty} \varsigma_k \;<\; \varsigma$$

then $\limsup_{k\to\infty} \| \tilde{J}^{\mu_k} - \tilde{J}^* \| < \epsilon$ *and* $\limsup_{k\to\infty} \| \tilde{J}^{\delta_k} - \tilde{J}^* \| < \epsilon$.

2. *Given any scalar $\epsilon > 0$, there exists a number $\varsigma > 0$ and a positive vector $\bar{\epsilon} \in \mathbf{R}^n$, with $\underline{0} < \bar{\epsilon} < \tilde{\epsilon}$, such that, if $\bar{\epsilon}_k < \bar{\epsilon}$ and $\varsigma_k < \varsigma$ for all k, then $\tilde{J}^{\mu_k}$*

converges to $\tilde{J}^*$ *in a finite number of steps* ($\leq |\Upsilon|$) *and* $\| \tilde{J}^{\delta_k} - \tilde{J}^{\mu_k} \| < \epsilon$ *for all* k.

3. *In particular if* $\limsup_{k\to\infty} \bar{\epsilon}_k(i) = 0, \ \forall i \in \{1, 2, \ldots, n\}$ *and* $\limsup_{k\to\infty} \varsigma_k = 0$, *then* $\| \tilde{J}^{\mu_k} - \tilde{J}^* \| \overset{k\to\infty}{\longrightarrow} 0$ *and* $\| \tilde{J}^{\delta_k} - \tilde{J}^* \| \overset{k\to\infty}{\longrightarrow} 0$.

□

Here $\tilde{J}^*$ is the optimal cost to go vector for the SSP problem and $\| \cdot \|$ is the sup-norm.

Proof of Theorem D.2

Note that $\tilde{J}^\delta$ is a uniformly continuous function on Λ (see Section D.1). Also in view of Theorem D.1 (as well as the comments following it), the conclusions of Theorem D.2 hold.

□

Note that in Algorithm D.3, instead of using an extremal policy $\delta_k \in \Lambda_{\bar{\epsilon}_k}$ to approximate the μ_k, we could have chosen any $\delta_k \in \text{Interior}(\Lambda)$, such that $[\delta_k(i)]_{\mu_k(i)} \geq (1 - (|\mathcal{A}(i)| - 1)\bar{\epsilon}_k(i)), \ \forall i \in \{1, 2, \ldots, n\}$; for instance δ_k could be made to depend on the estimated Q values in the previous step, namely $\tilde{Q}_{k-1}(\cdot, \cdot)$, assigning for each state $i \in \{1, 2, \ldots, n\}$, smaller probabilities to actions with larger Q values.

Consider the following algorithm for estimating Q values, when we are given a fixed vector $J \in \mathbf{R}^n$ as the one step terminal cost. That is we are interested in estimating $Q^J(i, a) = g(i, a) + \sum_{j=1}^n p_{ij}(a) J(j)$ for $(i, a) \in \mathcal{Q}$. We assume Assump-

tion D.3 to hold.

Algorithm D.4

Input : One step terminal cost $J \in R^n$ *and Stationary Randomized Policy* $\delta \in$ *Interior*(Λ).

Output : Estimate $\tilde{Q}(i,a)$ *of* $Q^J(i,a) = g(i,a) + \sum_{j=1}^{n} p_{ij}(a)J(j)$, *for* $(i,a) \in \mathcal{Q}$.

t *: index of trajectory;* $t \in \{1, 2, \ldots\}$.

m *: index of stage within each trajectory;* $m \in \{0, 1, 2, \ldots\}$.

$\hat{n}$ *: number of trajectories simulated, a positive integer.*

N_t *: stopping time for* t^{th} *trajectory.*

i_m^t *: state at* m^{th} *stage of* t^{th} *trajectory.*

a_m^t *: action taken at* m^{th} *stage of the* t^{th} *trajectory, from state* i_m^t *under policy* δ.

$g_{m,t}$ *: immediate cost incurred at stage* m *of* t^{th} *trajectory, when action* a_m^t *is taken from state* i_m^t.

$Q_{t,m}(i,a)$ *: estimate of* $Q^J(i,a)$ *at stage* m *of* t^{th} *trajectory.*

$\tau_m^t(i,a)$ *: number of times action 'a' has been taken from state 'i', by the time (including stage* m*) stage* m *is reached in the* t^{th} *trajectory.*

1.

$$\begin{aligned} t &= 1 \\ \tau_{-1}^{1}(i,a) &= 0 \qquad \forall (i,a) \in \mathcal{Q} \end{aligned}$$

The initial values

$$Q_{1,0}(i,a) \qquad \textit{arbitrary} \;\; \forall (i,a) \in \mathcal{Q}$$

2. For $m = 0$ to $N_t - 1$, do

$$\begin{aligned} \tau_m^t(i_m^t, a_m^t) &= \tau_{m-1}^t(i_m^t, a_m^t) + 1 \\ \tau_m^t(i,a) &= \tau_{m-1}^t(i,a), \quad \forall (i,a) \neq (i_m^t, a_m^t), \;\; (i,a) \in \mathcal{Q} \\ Q_{t,m+1}(i_m^t, a_m^t) &= \left(1 - \gamma_{\tau_m^t(i_m^t, a_m^t)}(i_m^t, a_m^t)\right) Q_{t,m}(i_m^t, a_m^t) \\ &\quad + \gamma_{\tau_m^t(i_m^t, a_m^t)}(i_m^t, a_m^t) \left(g_{m,t} + J(i_{m+1}^t)\right) \qquad \text{(D.6)} \\ Q_{t,m+1}(i,a) &= Q_{t,m}(i,a), \quad \forall (i,a) \neq (i_m^t, a_m^t), \quad (i,a) \in \mathcal{Q} \end{aligned}$$

3.

$$\begin{aligned} Q_{t+1,0}(i,a) &= Q_{t,N_t}(i,a) \qquad \forall (i,a) \in \mathcal{Q} \\ \tau_{-1}^{t+1}(i,a) &= \tau_{N_t-1}^t(i,a) \qquad \forall (i,a) \in \mathcal{Q} \end{aligned}$$

4. $t = t + 1$.

5. go to step 2, if $t \leq \hat{n}$; else go to step 6.

6. Return

$$\tilde{Q}(i,a) = Q_{t,0}(i,a) \qquad \forall (i,a) \in \mathcal{Q} \qquad \square$$

Note that '$J(0)$' is defined as zero. Now each starting state of the t^{th} trajectory $i_0^t \in \{1, 2, \ldots, n\}$ may be chosen based on the past history until that time.

We impose the condition that each state in $\{1, 2, \ldots, n\}$ is visited infinitely often (i.e. actions are taken from every state in $\{1, 2, \ldots, n\}$ infinitely often) if an infinite number of trajectories are generated in Algorithm D.4. Since $\delta \in \text{Interior}(\Lambda)$, each state action pair $(i, a) \in \mathcal{Q}$ will be taken infinitely often too.

For example we might choose a probability distribution (concentrated on a subset of $\{1, 2, \ldots, n\}$) on the initial state of each trajectory that is identical and independent of the past. Also the stopping time N_t may be taken to be the first time the t^{th} trajectory reaches state '0'. We assume that there is a positive probability of reaching any state from the starting states under policy $\delta \in \text{Interior}(\Lambda)$.

The comments following Algorithm D.1 on the nature of $\gamma_k(i, a)$ are also valid for Algorithm D.4.

Since each state action pair $(i, a) \in \mathcal{Q}$ is taken infinitely often, standard results from stochastic approximation theory [16] can be used to show that under Assumption D.3 we have $Q_{t,0}(i, a) \stackrel{t \to \infty}{\longrightarrow} Q^J(i, a)$ for $(i, a) \in \mathcal{Q}$.

Consider the following algorithm.

Algorithm D.5 *Let $\bar{\epsilon}_k$ be a sequence of positive vectors in $\mathbf{R}^n$ such that $\underline{0} < \bar{\epsilon}_k < \tilde{\epsilon}$, where the inequality is componentwise. Here $\tilde{\epsilon}(i) = \frac{1}{|\mathcal{A}(i)|}$, for $i \in \{1, 2, \ldots, n\}$.*

1. *Set $k = 0$.*

2. *Select an arbitrary stationary deterministic policy $\mu_0 \in \Upsilon$.*

3. *With policy μ_k run the off-line TD algorithm in Section D.3 (or alternatively the on-line TD algorithm in Section D.4) for a sufficiently large random number of trajectories, say $\tilde{n}_k$, till the cost to go estimates "nearly converges" to the actual cost to go vector $\tilde{J}^{\mu_k}$. Let $\tilde{J}_k$ be the estimate obtained at the end of this TD scheme.*

4. *Choose the stationary randomized extremal policy $\delta_k \in \Lambda_{\bar{\epsilon}_k}$ associated with policy μ_k. Run the Algorithm D.4 with one step terminal cost $\tilde{J}_k$ using policy δ_k for a sufficiently large random number of trajectories, say $\hat{n}_k$ till the Q values "nearly converges" to $Q^{\tilde{J}_k}(i,a) = g(i,a) + \sum_{j=1}^{n} p_{ij}(a)\, \tilde{J}_k(j),\ \forall (i,a) \in \mathcal{Q}$. Let $\tilde{Q}_k(i,a),\ \forall (i,a) \in \mathcal{Q}$ be the estimate of the Q values obtained at the end of Algorithm D.4.*

   Let

$$\varsigma_k = \max \left\{ \max_{i \in \{1,2,\ldots,n\}} | \left(\tilde{J}_k(i) - \tilde{J}^{\mu_k}(i) \right) |, \max_{(i,a) \in \mathcal{Q}} |\tilde{Q}_k(i,a) - Q^{\tilde{J}_k}(i,a)| \right\}$$

5. *Set $k = k+1$ and update the policy μ_k, where*

$$\mu_k(i) = \arg \min_{a \in \mathcal{A}(i)} \tilde{Q}_{k-1}(i,a), \qquad \forall i \in \{1,2,\ldots,n\}$$

6. *Go to step 3.*

□

Note that we stick with either the off-line TD scheme or the on-line TD scheme in step 3, throughout the execution of Algorithm D.5.

The number of trajectories simulated, namely $\tilde{n}_k$, inside the invocation of the TD scheme in step 3, of Algorithm D.5 may be decided inside the TD scheme for sufficiently "close convergence" of the cost to go estimate. Similarly the number of trajectories simulated, namely $\hat{n}_k$, inside the invocation of Algorithm D.4, in step 4 of Algorithm D.5 may be decided inside Algorithm D.4 for sufficiently "close convergence" of the Q values.

Though the initial estimate of the cost to go vector when calling the TD scheme in step 3 of Algorithm D.5 may be arbitrary, we may set it to the final estimate obtained in the previous iteration (i.e. $\tilde{J}_{k-1}$).

Similarly the initial estimates of the Q values when calling Algorithm D.4 in step 4 of Algorithm D.5 may be arbitrary, but can be set to the final estimates obtained in the previous iteration (i.e. $\tilde{Q}_{k-1}(\cdot,\cdot)$).

We have the following theorem which is similar in spirit to Theorem D.2.

Theorem D.3 *Consider Algorithm D.5 on an SSP problem where all stationary deterministic policies are proper. Assume that Assumption D.1, Assumption D.2 and Assumption D.3 hold.*

1. *Given any scalar $\epsilon > 0$, there exists an $\bar{\epsilon} \in \mathbf{R}^n$ with $\underline{0} < \bar{\epsilon} < \tilde{\epsilon}$ and a number $\varsigma > 0$, such that if*

$$\limsup_{k\to\infty} \bar{\epsilon}_k(i) \;<\; \bar{\epsilon}(i), \qquad \forall i \in \{1,2,\ldots,n\}$$

and

$$\limsup_{k\to\infty} \varsigma_k \;<\; \varsigma$$

then $\limsup_{k\to\infty} \| \tilde{J}^{\mu_k} - \tilde{J}^* \| < \epsilon$ *and* $\limsup_{k\to\infty} \| \tilde{J}^{\delta_k} - \tilde{J}^* \| < \epsilon$.

2. *There exists a scalar* $\varsigma > 0$ *such that if* $\varsigma_k < \varsigma$ *for all* k*, then* $\tilde{J}^{\mu_k}$ *converges to* $\tilde{J}^*$ *in a finite number of steps* ($\leq |\Upsilon|$)*. Furthermore given any scalar* $\epsilon > 0$*, there exists a positive vector* $\bar{\epsilon} \in \mathbf{R}^n$*, with* $\underline{0} < \bar{\epsilon} < \tilde{\epsilon}$*, such that, if* $\bar{\epsilon}_k < \bar{\epsilon}$*, then* $\| \tilde{J}^{\delta_k} - \tilde{J}_{\mu_k} \| < \epsilon$ *for all* k*.*

3. *In particular if* $\limsup_{k\to\infty} \bar{\epsilon}_k(i) = 0, \ \forall i \in \{1, 2, \ldots, n\}$ *and* $\limsup_{k\to\infty} \varsigma_k = 0$*, then* $\| \tilde{J}^{\mu_k} - \tilde{J}^* \| \overset{k\to\infty}{\longrightarrow} 0$ *and* $\| \tilde{J}^{\delta_k} - \tilde{J}^* \| \overset{k\to\infty}{\longrightarrow} 0$*.*

□

Here $\tilde{J}^*$ is the optimal cost to go vector for the SSP problem and $\| \cdot \|$ is the sup-norm.

The comments about the choice of δ_k following Theorem D.2 applies for the choice of δ_k in Algorithm D.5 too.

With the equivalent SSP formulation of discounted cost problem, we can solve the discounted cost optimal cost problem using the above algorithms.

We may also solve the infinite horizon discounted cost problem with discount factor β ($0 \leq \beta < 1$), directly, with slight variants of Algorithm D.1, Algorithm D.2, Algorithm D.3, Algorithm D.4 and Algorithm D.5.

See Section D.6 for the variant of the TD schemes for discounted cost problems, where we try to estimate for a policy $\delta \in \Lambda$, the discounted cost to go vector $J^\delta = (I - \beta P_\delta)^{-1} \bar{g}^\delta$. Here P_δ is the stochastic transition matrix corresponding to policy δ, and $\bar{g}^\delta$ is the immediate cost vector corresponding to policy δ. Note the variations in the requirements of the eligibility coefficients mentioned in Section D.6.

In the offline-line TD scheme (Section D.3) we have the modified temporal difference

$$d_{m,t} = g_{m,t} + \beta\, J_t(i^t_{m+1}) - J_t(i^t_m)$$

whereas in the on-line TD scheme (Section D.4) we have the modified temporal difference

$$d^0_{m,t} = g_{m,t} + \beta\, J^0_{t,m}(i^t_{m+1}) - J^0_{t,m}(i^t_m)$$

In the variant of Algorithm D.1 and Algorithm D.2 we try to estimate the discounted cost to go vector J^δ and the corresponding Q values given by $Q^\delta(i,a) = g(i,a) + \beta \sum_{j=1}^n p_{ij}(a)\, J^\delta(j)$.

Correspondingly we make the following modifications in Algorithm D.1. First we modify

$$d_{m,t} = g_{m,t} + \beta\, J_t(i^t_{m+1}) - J_t(i^t_m)$$

Also in equation D.4 we make the following modification,

$$\begin{aligned} Q_{t,m+1}(i^t_m, a^t_m) &= \left(1 - \gamma_{\tau^t_m(i^t_m,a^t_m)}(i^t_m, a^t_m)\right) Q_{t,m}(i^t_m, a^t_m) \\ &\quad + \gamma_{\tau^t_m(i^t_m,a^t_m)}(i^t_m, a^t_m)\left(g_{m,t} + \beta\, J_t(i^t_{m+1})\right) \end{aligned}$$

Similarly we make the following modifications in Algorithm D.2. First we modify,

$$d^0_{m,t} = g_{m,t} + \beta\, J^0_{t,m}(i^t_{m+1}) - J^0_{t,m}(i^t_m)$$

Also in equation D.5 we make the following modification,

$$\begin{aligned} Q_{t,m+1}(i^t_m, a^t_m) &= \left(1 - \gamma_{\tau^t_m(i^t_m,a^t_m)}(i^t_m, a^t_m)\right) Q_{t,m}(i^t_m, a^t_m) \\ &\quad + \gamma_{\tau^t_m(i^t_m,a^t_m)}(i^t_m, a^t_m)\left(g_{m,t} + \beta\, J^0_{t,m+1}(i^t_{m+1})\right) \end{aligned}$$

In the variant of Algorithm D.3, in step 3 we try to estimate the discounted cost to go vector J^{δ_k} and the Q values given by $Q^{\delta_k}(i,a) = g(i,a) + \beta \sum_{j=1}^{n} p_{ij}(a)\, J^{\delta}(j)$ for $(i,a) \in \mathcal{Q}$.

In the variant of Algorithm D.4 we try to estimate

$$Q^{J}(i,a) = g(i,a) + \beta \sum_{j=1}^{n} p_{ij}(a) J(j)$$

Correspondingly in equation D.6 in Algorithm D.4 we make the following modification,

$$\begin{aligned} Q_{t,m+1}(i_m^t, a_m^t) &= \left(1 - \gamma_{\tau_m^t(i_m^t,a_m^t)}(i_m^t, a_m^t)\right) Q_{t,m}(i_m^t, a_m^t) \\ &\quad + \gamma_{\tau_m^t(i_m^t,a_m^t)}(i_m^t, a_m^t) \left(g_{m,t} + \beta\, J(i_{m+1}^t)\right) \end{aligned}$$

Similarly in the variant of Algorithm D.5, in step 3, we get $\tilde{J}_k$, the estimate of the discounted cost to go vector J^{μ_k}, for policy μ_k. In step 4 we try to estimate $Q^{\tilde{J}_k}(i,a) = g(i,a) + \beta \sum_{j=1}^{n} p_{ij}(a) \tilde{J}_k(j)$ by calling the variant of Algorithm D.4.

As an aside, see [16, Section 6.3] for TD(λ) schemes with linear function approximation (instead of lookup table schemes) to approximate the cost to go function for SSP problems for a fixed stationary policy.

Bibliography

[1] Abounadi J., Bertsekas D. and Borkar V. S., Learning Algorithms for Markov Decision Processes with Average Cost, *SIAM Journal of Control and Optimization*, Vol. 40, No. 3, (2001), 681-698.

[2] Abounadi J., Bertsekas D. and Borkar V. S., Stochastic Aprroximation for Nonexpansive Maps: Application to Q-Learning Algorithms, *SIAM Journal of Control and Optimization*, Vol. 41, No. 1, (2002), 1-22.

[3] Apostol T. M., Mathematical Analysis, Addison-Wesley/Narosa, 1974.

[4] Arapostathis A., Borkar V. S., Fernández Gaucherand E., Ghosh M. K. and Marcus S. I., Discrete-Time Controlled Markov Processes with Average Cost Criterion: A Survey, *SIAM Journal of Control and Optimization*, Vol. 31, No. 2, March (1993), 282-344.

[5] Åström K. J., Optimal Control of Markov Processes with Incomplete State Information, *Journal of Mathematical Analysis and Applications*, 10, (1965), 174-205.

[6] Åström K. J., Optimal Control of Markov Processes with Incomplete State Information. II. The Convexity of the Lossfunction, *Journal of Mathematical Analysis and Applications*, 26, (1969), 403-406.

[7] Barto A. G., Bradtke S. J. and Singh S. P., Learning to Act using Real-Time Dynamic Programming, *Artificial Intelligence*, 72, (1995), 81-138.

[8] Benveniste A., Metivier M. and Priouret P., Adaptive Algorithms and Stochastic Approximations, Springer-Verlag, 1990.

[9] Bertsekas D. P., Convergence of Discretization Procedures in Dynamic Programming, *IEEE Transactions on Automatic Control*, 20, (1975), 415-419.

[10] Bertsekas D. P., Dynamic Programming : Deterministic and Stochastic Models, Prentice-Hall, Englewood Cliffs, NJ, 1987.

[11] Bertsekas D. P., Dynamic Programming and Optimal Control Vol. 1, Third Edition, Athena Scientific, 2005.

[12] Bertsekas D. P., Dynamic Programming and Optimal Control Vol. 2, Third Edition, Athena Scientific, 2007.

[13] Bertsekas D. P. and Castanon D. A., Adaptive Aggregation Methods for Infinite Horizon Dynamic Programming,*IEEE Transactions on Automatic Control*, Vol. 34, No. 6, June (1989), 589-598.

[14] Bertsekas D. P. and Shreve S. E., Stochastic Optimal Control : The Discrete-Time Case, Academic Press, 1978.

[15] Bertsekas D. P. and Tsitsiklis J. N., Parallel and Distributed Computation : Numerical Methods, Prentice Hall, Englewood Cliffs, NJ, 1989.

[16] Bertsekas D. P. and Tsitsiklis J. N., Neuro-Dynamic Programming, Athena Scientific, Belmont, Massachusetts, 1996.

[17] Borkar V. S., Stochastic Approximation with Two Time Scales, Systems and Control Letters, 29, (1997), 291-294.

[18] Chrisman L., Reinforcement Learning with Perceptual Aliasing: The Predictive Distinctions Approach, *Procceedings of the Tenth National Conference on Artificial Intelligence (AAAI Press)*, San Jose, CA, July 12-16 (1992), 183-188.

[19] Dayan P., The Convergence of TD(λ) for General λ, *Machine Learning*, 8, (1992), 341-362.

[20] Dayan P. and Sejnowski T. J., TD(λ) Converges with Probability 1, *Machine Learning*, 14, (1994), 295-301.

[21] Durrett R., Probability : Theory and Examples, Duxbury Press, 3rd edition, 2005.

[22] Fernandez Gaucherand E., Arapostathis A. and Marcus S. I., Analysis of an Adaptive Control Scheme for a Partially Observable Controlled Markov Chain, *IEEE Transactions on Automatic Control*, Vol. 38, No. 6, June (1993), 987-993.

[23] Hernández-Lerma O., Adaptive Markov Control Processes. Springer-Verlag, New-York, 1989.

[24] Hernández-Lerma O. and Marcus S. I., Discretization Procedures for Adaptive Markov Control Processes, *Journal of Mathematical Analysis and Applications*, Vol. 137, No. 2,(1989), 485-514.

[25] Horn R. A. and Johnson,. R., Matrix Analysis, Cambridge University Press, 1985.

[26] Jaakkola T., Jordan M. I. and Singh S. P., On the Convergence of Stochastic Iterative Dynamic Programming Algorithms, *Neural Computation*, 6, (1994), 1185-1201.

[27] Jaakkola T., Singh S. P. and Jordan M. I., Reinforcement Learning Algorithm for Partially Observable Markov Decision Problem, *Advances in Neural Information Processing Systems*, 7 (NIPS), (1995), 345-352.

[28] Kaelbling L. P., Littman M. L. and Cassandra A. R., Planning and Acting in Partially Observable Stochastic Domains, *Artificial Intelligence*, 101, (1998), 99-134.

[29] Konda V. R. and Borkar V. S., Actor-Critic-Type Learning Algorithms for Markov Decision Processes, *SIAM Journal of Control and Optimization*, Vol. 38, No. 1, (1999), 94-123.

[30] Konda V. R. and Tsitsiklis J. N., On Actor-Critic Algorithms, *SIAM Journal of Control and Optimization*, Vol. 42, No. 4, (2003), 1143-1166.

[31] Littman M. L., Algorithms for Sequential Decision Making, Ph.D. Dissertation, Department of Computer Science, Brown University, Providence, RI (1996).

[32] Loch J. and Singh S., Using Eligibility Traces to Find the Best Memoryless Policy in Partially Observable Markov Decision Processes, *Proceedings of the Fifteenth International Conference on Machine Learning (ICML)*, (1998), 323-331.

[33] Lovejoy W. S., On the Convexity of Policy Regions in Partially Observed Systems, Technical Note, Operations Research, Vol. 35, No. 4, July-August (1987), 619-621.

[34] Lovejoy W. S., An Approximate Algorithm, with Bounds, for Composite State Partially Observed Markov Decision Processes, *Proceedings of the* 29^{th} *IEEE Conference on Decision Control*, Honolulu, Hawaii, December (1990), 1344-1348.

[35] Lovejoy W. S., A Survey of Algorithms for Partially Observed Markov Decision Processes, *Annals of Operations Research*, 28,(1991), 47-66.

[36] Lovejoy W. S., Computationally Feasible Bounds for Partially Observed Markov Decision Processes, *Operations Research*, Vol. 39, No. 1, Jan-Feb (1991), 162-175.

[37] Luenberger D. G., Optimization by Vector Space Methods, Wiley Interscience, 1969.

[38] Mahadevan S., Average Reward Reinforcement Learning: Foundations, Algorithms, and Empirical Results, *Machine Learning* 22, (1996), 159-195.

[39] Monahan G. E., A Survey of Partially Observable Markov Decision Processes: Theory, Models, and Algorithms, *Management Science*, Vol. 28, No. 1, January (1982), 1-16.

[40] Puterman M. L., Markov Decision Processes : Discrete Stochastic Dynamic Programming, John Wiley and Sons Inc, 2005.

[41] Royden H. L., Real Analysis, Third Edition, Macmillan Publishing Company, 1988.

[42] Rummery G. A., Problem Solving with Reinforcement Learning, Ph.D. Dissertation, Cambridge University Engineering Department, Cambridge, England, July (1995).

[43] Seneta E., Non-negative Matrices and Markov Chains, Springer Verlag, 1981.

[44] Singh S. P., Learning to Solve Markovian Decision Processes, Ph.D. Dissertation, Department of Computer Science, University of Massachusetts, Amherst, MA (1994).

[45] Singh S. P., Jaakkola T. and Jordan M. I., Learning Without State-Estimation in Partially Observable Markovian Decision Processes, *Machine Learning: Proceedings of the Eleventh International Conference on Machine Learning (ICML)*, (1994), 284-292.

[46] Singh S. P. and Yee R. C., An Upper Bound on the Loss from Approximate Optimal-Value Functions, *Machine Learning*, 16, (1994), 227-233.

[47] Smallwood R. D. and Sondik E. J., The Optimal Control of Partially Observable Markov Processes over a Finite Horizon, *Operations Research*, Vol. 21, (1973), 1071-1088.

[48] Sondik E. J., The Optimal Control of Partially Observable Markov Processes, Ph.D. Dissertation, Department of Electrical Engineering, Stanford University, Stanford, CA (1971).

[49] Sondik E. J., The Optimal Control of Partially Observable Markov Processes over the Infinite Horizon: Discounted Costs, *Operations Research*, Vol. 26, No. 2, March-April (1978), 282-304.

[50] Sutton R., Learning to Predict by the Methods of Temporal Differences, *Machine Learning*, 3, (1988), 9-44.

[51] Tsitsiklis J. N., Asynchronous Stochastic Approximation and Q-Learning, *Machine Learning*, 16, (1994), 185-202.

[52] Tsitsiklis J. N., On the Convergence of Optimistic Policy Iteration, *Journal of Machine Learning Research*, 3, (2002), 59-72.

[53] Tsitsiklis J. N. and Van Roy B., Feature-Based Methods for Large Scale Dynamic Programming, *Machine Learning*, 22,(1996), 59-94.

[54] Tsitsiklis J. N. and Van Roy B., An Analysis of Temporal Difference Learning with Function Approximation, *IEEE Transactions on Automatic Control*, Vol. 42, No. 5, May (1997), 674-690.

[55] Tsitsiklis J. N. and Van Roy B., Average Cost Temporal Difference Learning, *Automatica*, 35, (1999), 1799-1808.

[56] Tsitsiklis J. N. and Van Roy B., On Average Versus Discounted Reward Temporal Difference Learning, *Machine Learning*, 49, (2002), 179-191.

[57] Watkins J. C. H. and Dayan P., Q-Learning, *Machine Learning*, 8, (1992), 279-292.