ROBUSTNESS TO APPROXIMATIONS AND MODEL LEARNING IN MDPs AND POMDPs

by

ALI DEV Ran KARA

A thesis submitted to the Department of Mathematics and Statistics in conformity with the requirements for the degree of Doctor of Philosophy

Queen’s University
Kingston, Ontario, Canada
June 2021

Copyright © Ali Devran Kara, 2021
Abstract

This thesis develops a topological approach to robustness to incorrect-incomplete models in MDPs and POMDPs, and applications of this study to approximations, model learning and reinforcement learning.

In the first part of the thesis, we study continuity properties of multi-stage stochastic control problems with respect to transition kernels, and applications of these results to the study of robustness of optimal control policies designed for incorrect models applied to a true system, and the convergence (or lack of it) of the error due to mismatch as the incorrect model approaches the true one. We analyze the problem under the expected infinite horizon discounted cost setup for fully and partially observed models and under the expected infinite horizon average cost setup for fully observed models. We show that continuity can be established under total variation convergence of the transition kernels under mild assumptions, and with further restrictions on the dynamics under weak and setwise convergence of the transition kernels. Using these continuity properties, we establish positive and negative convergence results on robustness. Compared to the existing literature, we obtain strictly refined robustness results that are applicable even when the incorrect models can be investigated under weak convergence and setwise convergence criteria (with respect to a true model). These entail implications on empirical learning in (data-driven) stochastic
control since often system models are learned through empirical training data where typically the weak convergence criterion applies but stronger convergence criteria do not. We view approximation of MDPs with finite models as a particular instance of robustness and we establish how, in particular, continuous weak convergence applies under finite model quantization of weakly continuous kernels.

In the second part of the thesis, we focus on partially observed Markov decision processes (POMDPs). We first establish conditions on the weak Feller property of non-linear filter kernels for the study of POMDPs. Establishing sufficient conditions for the weak Feller property for such processes is a significant problem, studied under various assumptions and setups in the literature. We provide a concise and unified proof for the weak Feller property of the non-linear filter process (i) first under weak continuity of the transition probability of controlled Markov chain and total variation continuity of its observation channel (a result due to Feinberg et. al. [Math. Oper. Res. 41(2) (2016) 656-681]), and then, (ii) under total variation continuity of the transition probability of controlled Markov chain.

Using the obtained regularity properties, we then study approximation techniques for POMDPs and application of these approximations to Q learning algorithms for partially observed models using only a finite memory of information variables. We rigorously establish near optimality of finite window feedback control policies in POMDPs under mild non-linear filter stability conditions and the assumption that the measurement and action sets are finite (and the state space is real vector valued). We also establish a rate of convergence result which relates the finite window memory size and the approximation error bound. This rate of convergence is exponential under geometric filter stability conditions. We apply these results to Q learning algorithms for POMDPs that use a finite window of
history. We establish near optimality of policies that are learned using finite window history variables.

Finally, as a practically consequential application, we consider the Q learning problem for continuous state spaces and show near optimality of algorithms constructed by discretizing the state space where we view the problem as a partially observed model with a quantizer channel.
Acknowledgments

First, I would like to thank my supervisor Serdar Yuksel. It has been a privilege to work with such a great scholar and to learn how to pursue research from him. I will always be proud to say that I was his PhD student. Not only has he helped me with my research, he has also helped me with all aspects of life in that he has been more like a friend than a supervisor.

I thank my parents Battal and Meryem Kara for their unconditional love. Knowing you will always be there for me has given me strength during my studies and during my difficult times; I will always try my best to make you proud. I am also so happy to realize that as I get older I become more like you guys.

I am also thankful to my sister Ekin, she has been my biggest supporter. I am so lucky that my sister is also my best friend.

I would like to thank Cetin Tastekin. We met in 2012 and learned the university life and academics together, even now there are two continents between us, he is still helping me with the most difficult decisions. I will always be waiting for the songs you send me.

I met some great people and lifelong friends in Kingston. I want to thank Arda, Aybike and Onur for making Kingston a home for me, thanks to you guys, it never felt like "gurbet" here. I thank Bora, Fauz and Jake for being my Canadian mentors; they all taught me a different side of Canada, I will always have an excuse to come back to Canada thanks to
them.

I would like thank Jennifer Read; she has helped me with almost everything since even before I came to Canada. Thanks to her, I never had to follow what I had to do for my degree, I do not think I filled even a single form myself, she did it all for me. I can confidently say that I could not have completed my degree without you Jennifer.

Finally, I would like to thank Queen’s ARC facilities for keeping me going during my PhD.
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Abstract</strong></td>
<td>i</td>
</tr>
<tr>
<td></td>
<td><strong>Acknowledgments</strong></td>
<td>iv</td>
</tr>
<tr>
<td></td>
<td><strong>Contents</strong></td>
<td>vi</td>
</tr>
<tr>
<td></td>
<td><strong>List of Figures</strong></td>
<td>x</td>
</tr>
<tr>
<td></td>
<td><strong>Chapter 1: Introduction, Literature Review and Summary of Contributions</strong></td>
<td></td>
</tr>
<tr>
<td>1.1</td>
<td>Model Description</td>
<td>1</td>
</tr>
<tr>
<td>1.2</td>
<td>A General Literature Review and Our Contributions</td>
<td>5</td>
</tr>
<tr>
<td>1.3</td>
<td>Outline of the Thesis</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td><strong>Chapter 2: Continuity and Robustness under Incorrect Transition Kernels under Infinite Horizon Discounted Cost Criterion</strong></td>
<td>17</td>
</tr>
<tr>
<td>2.1</td>
<td>Literature Review</td>
<td>18</td>
</tr>
<tr>
<td>2.2</td>
<td>Some Examples and Convergence Criteria for Transition Kernels</td>
<td>21</td>
</tr>
<tr>
<td>2.2.1</td>
<td>Convergence Criteria for Transition Kernels</td>
<td>21</td>
</tr>
<tr>
<td>2.2.2</td>
<td>Examples</td>
<td>23</td>
</tr>
<tr>
<td>2.3</td>
<td>Summary of Results and Contributions of the Chapter</td>
<td>26</td>
</tr>
<tr>
<td>2.4</td>
<td>Continuity of Optimal Cost with respect to Convergence of Transition Kernels (Partially Observed Case)</td>
<td>29</td>
</tr>
<tr>
<td>2.4.1</td>
<td>Weak Convergence</td>
<td>29</td>
</tr>
<tr>
<td>2.4.2</td>
<td>A Sufficient Condition for Continuity under Weak Convergence</td>
<td>31</td>
</tr>
<tr>
<td>2.4.3</td>
<td>Absence of Continuity under Setwise Convergence</td>
<td>34</td>
</tr>
<tr>
<td>2.4.4</td>
<td>Continuity under Total Variation</td>
<td>35</td>
</tr>
<tr>
<td>2.5</td>
<td>Robustness to Incorrect Transition Kernels (Partially Observed Case)</td>
<td>38</td>
</tr>
<tr>
<td>2.5.1</td>
<td>Total Variation</td>
<td>38</td>
</tr>
<tr>
<td>2.5.2</td>
<td>Setwise Convergence</td>
<td>40</td>
</tr>
<tr>
<td>2.5.3</td>
<td>Weak Convergence</td>
<td>40</td>
</tr>
<tr>
<td>2.6</td>
<td>Continuity and Robustness in the Fully Observed Case</td>
<td>42</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
<td></td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>2.6.1 Weak Convergence</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>2.6.2 Setwise Convergence</td>
<td>54</td>
<td></td>
</tr>
<tr>
<td>2.6.3 Total Variation</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>2.7 Finite Model Approximations for Weakly Continuous MDPs as an Instance of the Robustness Problem</td>
<td>61</td>
<td></td>
</tr>
<tr>
<td>2.7.1 Review of Finitely Quantized Approximations to Standard Borel MDPs</td>
<td>61</td>
<td></td>
</tr>
<tr>
<td>2.8 Empirical Model Learning as an Instance of the Robustness Problem</td>
<td>66</td>
<td></td>
</tr>
<tr>
<td>2.9 Technical Results</td>
<td>73</td>
<td></td>
</tr>
<tr>
<td>2.9.1 Proof of (2.3) in Theorem 2.4.2</td>
<td>73</td>
<td></td>
</tr>
<tr>
<td>2.9.2 Proof of (2.5) in Theorem 2.4.5</td>
<td>79</td>
<td></td>
</tr>
<tr>
<td>2.9.3 Proof of (2.7) in Theorem 2.4.6</td>
<td>82</td>
<td></td>
</tr>
<tr>
<td>2.9.4 Proof for Remark 2.6.3</td>
<td>84</td>
<td></td>
</tr>
<tr>
<td>Chapter 3: Continuity and Robustness under Incorrect Transition Kernels under Infinite Horizon Average Cost Criterion</td>
<td>88</td>
<td></td>
</tr>
<tr>
<td>3.1 Literature Review and Contributions</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td>3.2 Convergence Criteria for Transition Kernels</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>3.3 Supporting Results: Continuity under Convergence of Transition Kernels</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>3.3.1 Some Differences with the Infinite Horizon Discounted Problem</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>3.3.2 Ergodicity Properties of Controlled Markov Chains</td>
<td>93</td>
<td></td>
</tr>
<tr>
<td>3.3.3 Optimality of Stationary Policies</td>
<td>94</td>
<td></td>
</tr>
<tr>
<td>3.3.4 Approximation by Finite Horizon Cost</td>
<td>95</td>
<td></td>
</tr>
<tr>
<td>3.3.5 Continuity under the Convergence of Transition Kernels</td>
<td>97</td>
<td></td>
</tr>
<tr>
<td>3.4 Robustness to Incorrect Controlled Transition Kernel Models</td>
<td>98</td>
<td></td>
</tr>
<tr>
<td>3.4.1 Robustness under Weak Convergence of Transition Kernels</td>
<td>99</td>
<td></td>
</tr>
<tr>
<td>3.4.2 Robustness under Setwise Convergence of Transition Kernels</td>
<td>104</td>
<td></td>
</tr>
<tr>
<td>3.4.3 Robustness under Total Variation Convergence of Transition Kernels</td>
<td>106</td>
<td></td>
</tr>
<tr>
<td>3.5 Comparison with the Literature</td>
<td>109</td>
<td></td>
</tr>
<tr>
<td>3.6 Robustness under Convergence of Transition Kernels without Uniform Ergodicity</td>
<td>111</td>
<td></td>
</tr>
<tr>
<td>3.7 Application to Empirical Model Learning</td>
<td>114</td>
<td></td>
</tr>
<tr>
<td>3.7.1 Empirical Model Learning and Robustness</td>
<td>114</td>
<td></td>
</tr>
<tr>
<td>3.7.2 Adaptive Learning</td>
<td>117</td>
<td></td>
</tr>
<tr>
<td>3.8 Technical Results</td>
<td>118</td>
<td></td>
</tr>
<tr>
<td>3.8.1 Proof of Equation (3.7)</td>
<td>118</td>
<td></td>
</tr>
<tr>
<td>3.8.2 Ergodicity conditions on Markov Chains</td>
<td>120</td>
<td></td>
</tr>
<tr>
<td>Chapter 4: Regularity Relations Between POMDPs and Their Fully Observed Belief MDP Reduction</td>
<td>122</td>
<td></td>
</tr>
</tbody>
</table>
## Chapter 4: Main Results and Connections with the Literature

4.1 Markov Property of Filter Processes ................................................. 122
4.2 Problem Formulation ...................................................................... 126
4.3 Significance of the Problem ............................................................ 127
4.4 Main Results and Connections with the Literature ......................... 129
   4.4.1 Statement of Main Results ..................................................... 129
   4.4.2 Examples ............................................................................... 131
   4.4.3 Comparison with the Prior Results ........................................ 133
   4.4.4 Proofs ................................................................................. 137
   4.4.5 A Technical Generalization .................................................. 152
4.5 Conclusion ...................................................................................... 155

## Chapter 5: Approximations in Partially Observed Models Using Finite Memory

5.1 Literature Review ............................................................................ 157
5.2 Regularity and Stability Properties of the Belief-MDP ................... 160
   5.2.1 Ergodicity and Filter Stability Properties of Partially Observed MDPs 161
5.3 Method I: Belief-space Quantization based on Finite-Memory .......... 163
   5.3.1 Approximation Error Analysis and Rates of Convergence ............ 167
5.4 Method II: Finite Window Memory as Approximate POMDP State ...... 178
   5.4.1 An Alternative Finite Window Belief-MDP Reduction .................. 178
   5.4.2 Approximation of the Finite Window Belief-MDP ....................... 183
   5.4.3 Approximation Error Analysis and Rates of Convergence ............ 185
5.5 Comparison of Approximation Methods ......................................... 188
5.6 Numerical Study ............................................................................. 190
5.7 Proofs of Technical Results ............................................................ 194
   5.7.1 Proof of Theorem 5.3.1 .......................................................... 194
   5.7.2 Proof of Step 1 in the Proof of Lemma 5.7.1 ............................... 205
   5.7.3 Proof of Step 2 in the Proof of Lemma 5.7.1 ............................... 207
   5.7.4 Proof of Lemma 5.7.5 .............................................................. 210
   5.7.5 Proof of Lemma 5.7.6 .............................................................. 215
   5.7.6 Proof of Theorem 5.4.1 ............................................................ 216
   5.7.7 Proof of Theorem 5.4.2 ............................................................ 219

## Chapter 6: Convergence of Finite Memory based Learning and Near Optimality under Filter Stability

6.1 Q Learning Algorithm for Fully Observed MDPs ............................ 223
6.2 Literature Review ............................................................................ 225
6.3 Approximate Belief MDP Construction with Finite Memory Feedback Revisited ................................................................. 229
   6.3.1 An Alternative Finite Window Belief-MDP Reduction ................. 229
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.3.2 Approximation of the Finite Window Belief-MDP</td>
<td>233</td>
</tr>
<tr>
<td>6.4 Q Iterations Using a Finite History of Information Variables and Convergence</td>
<td>235</td>
</tr>
<tr>
<td>6.5 Convergence to Near Optimality under Filter Stability</td>
<td>246</td>
</tr>
<tr>
<td>6.6 Quantization as a Measurement Channel: Q-Learning for Continuous Space MDPs via Quantization and Near Optimality</td>
<td>249</td>
</tr>
<tr>
<td>6.6.1 Construction of a Finite MDP</td>
<td>250</td>
</tr>
<tr>
<td>6.6.2 Approximate Q Learning for Continuous State Spaces</td>
<td>252</td>
</tr>
<tr>
<td>6.7 Numerical Study</td>
<td>254</td>
</tr>
<tr>
<td>6.8 Technical Results</td>
<td>257</td>
</tr>
<tr>
<td>Chapter 7: Conclusion and Future Problems</td>
<td>262</td>
</tr>
<tr>
<td>Bibliography</td>
<td>265</td>
</tr>
</tbody>
</table>
List of Figures

5.1 Approximation error in memory size for different methods. . . . . . . . . . 192
5.2 Approximation error, BL filter stability and Dobrushin coefficient in mem-
ory size for the first method. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 193
5.3 Approximation error, TV filter stability and Dobrushin coefficient in mem-
ory size for the second method. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 193
5.4 Overall comparison . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 194
6.1 Approximation error in memory size for different methods. . . . . . . . . 256
6.2 Approximation error in memory size for different methods. . . . . . . . . 257
Chapter 1

Introduction, Literature Review and Summary of Contributions

Control theory deals with influencing the behavior of a dynamical system to achieve a desired performance objective. The performance of a control policy heavily depends on how well the assumed model reflects the real world dynamics of the system. Often, the designer lacks the perfect knowledge of the dynamical system, and in many other cases, the perfect model is not feasible to work with, either analytically or computationally. Lack of perfect knowledge of the system and/or the need for approximations bring up the study of controllers that are robust to uncertainties and approximations. In what follows, we will describe the model that we will be investigating the robustness problem for.

1.1 Model Description

In this thesis, our focus will be on discrete time controlled stochastic processes. We will study fully observed and partially observed settings. If the decision maker (DM) has direct access to the state process, we will call the system a fully observed model, if the DM observes the state through a measurement channel, then we will call the system a partially
1.1. MODEL DESCRIPTION

observed model. To keep the introductory model as general as possible, we present the mathematical model for a partially observed Markov decision process.

Let $X \subset \mathbb{R}^m$ denote a Borel set which is the state space of a controlled Markov process. Here and throughout the thesis $\mathbb{Z}_+$ denotes the set of non-negative integers and $\mathbb{N}$ denotes the set of positive integers. Let $Y \subset \mathbb{R}^n$ be a Borel set denoting the observation space of the model, and let the state be observed through an observation channel $Q$. The observation channel, $Q$, is defined as a stochastic kernel (regular conditional probability) from $X$ to $Y$, such that $Q(\cdot|x)$ is a probability measure on the (Borel) $\sigma$-algebra $\mathcal{B}(Y)$ of $Y$ for every $x \in X$, and $Q(A|\cdot) : X \to [0,1]$ is a Borel measurable function for every $A \in \mathcal{B}(Y)$.

A decision maker (DM) is located at the output of the channel $Q$, hence, it only sees the observations $\{Y_t, t \in \mathbb{Z}_+\}$ and chooses its actions from $U \subset \mathbb{R}^l$, the action space which is a Borel subset of some Euclidean space. An admissible policy $\gamma$ is a sequence of control functions $\{\gamma_t, t \in \mathbb{Z}_+\}$ such that $\gamma_t$ is measurable with respect to the $\sigma$-algebra generated by the information variables

$$I_t = \{Y_{[0,t]}, U_{[0,t-1]}\}, \quad t \in \mathbb{N}, \quad I_0 = \{Y_0\},$$

where

$$U_t = \gamma_t(I_t), \quad t \in \mathbb{Z}_+ \tag{1.1}$$

are the $U$-valued control actions and

$$Y_{[0,t]} = \{Y_s, 0 \leq s \leq t\}, \quad U_{[0,t-1]} = \{U_s, 0 \leq s \leq t-1\}.$$

We define $\Gamma$ to be the set of all such admissible policies.
1.1. MODEL DESCRIPTION

The update rules of the system are determined by (1.1) and the following relationships:

\[
\Pr((X_0, Y_0) \in B) = \int_B P(dx_0)Q(dy_0|x_0), \quad B \in \mathcal{B}(X \times Y),
\]

where \(P\) is the (prior) distribution of the initial state \(X_0\), and

\[
\Pr\left((X_t, Y_t) \in B \middle| (X, Y, U)_{[0,t-1]} = (x, y, u)_{[0,t-1]}\right) = \int_B T(dx_t|x_{t-1}, u_{t-1})Q(dy_t|x_t), \quad B \in \mathcal{B}(X \times Y), \quad t \in \mathbb{N},
\]

where \(T\) is the transition kernel of the model which is a stochastic kernel from \(X \times U\) to \(X\).

Using stochastic realization results (see Lemma 1.2 in \[43\], or Lemma 3.1 of \[17\]), the process defined above can equivalently be represented in a functional form as follows:

\[
X_{t+1} = f(X_t, U_t, W_t), \quad Y_t = g(X_t, V_t) \quad (1.2)
\]

for some measurable functions \(f\) and \(g\), with \(\{W_t\}\) being an independent and identically distributed (i.i.d.) system noise process and \(\{V_t\}\) an i.i.d. disturbance process, which are independent of \(X_0\) and each other. Here, the first equation represents the transition kernel \(T\) as it gives the relation of the most recent state and action variables to the upcoming state. With this representation, it can be seen that the probabilistic nature of the transition kernel is determined by the function \(f\) and the probability model of the noise \(W_t\). The second equation represents the measurement channel \(Q\), as it describes the relation between the state and observation variables.

Throughout the thesis, we study two different cost criteria: infinite horizon expected
discounted cost and infinite horizon expected average cost. We let the objective of the 
agent (decision maker) be the minimization of these cost functions:

\[ J_\beta(P, \mathcal{T}, \gamma) = E_{P}^{\mathcal{T}, \gamma} \left[ \sum_{t=0}^{\infty} \beta^t c(X_t, U_t) \right], \quad \beta \in (0, 1), \]

\[ J_\infty(P, \mathcal{T}, \gamma) = \limsup_{N \to \infty} \frac{1}{N} \sum_{t=0}^{N-1} E_{P}^{\mathcal{T}, \gamma} [c(X_t, U_t)], \]

over the set of admissible policies \( \gamma \in \Gamma \), where \( c : X \times U \to \mathbb{R} \) is a Borel-measurable 
stage-wise cost function and \( E_{P}^{\mathcal{T}, \gamma} \) denotes the expectation with initial state probability 
measure \( P \) and transition kernel \( \mathcal{T} \) under policy \( \gamma \). Note that \( P \in \mathcal{P}(X) \), where we let 
\( \mathcal{P}(X) \) denote the set of probability measures on \( X \).

We define the optimal costs as functions of the priors and the transition kernels as

\[ J_\beta^*(P, \mathcal{T}) = \inf_{\gamma \in \Gamma} J_\beta(P, \mathcal{T}, \gamma), \]

\[ J_\infty^*(P, \mathcal{T}) = \inf_{\gamma \in \Gamma} J_\infty(P, \mathcal{T}, \gamma). \]

Note that the infinite horizon expected discounted cost and infinite horizon expected aver-
age cost depend on the initial distribution \( P \), transition kernel \( \mathcal{T} \), control policy \( \gamma \) (and the 
channel \( Q \) if the state is observed through a channel) since they determine the probability 
measure for the expectation operator. Here and often in the thesis, we will write the cost 
as a function of \( P \) and \( \mathcal{T} \), since we will be interested in the effects of transition kernels on 
the optimal cost functions. However, in the following chapters, we might sometimes alter 
the notation and write the costs as functions of other components of the model such as the 
channel \( Q \) or the initial state \( x_0 \) if the state is directly observable. These changes on the 
notation, hopefully, will be clear from the context.
1.2 A GENERAL LITERATURE REVIEW AND OUR CONTRIBUTIONS

Robustness is a desired property for optimal control of stochastic or deterministic systems when a given model does not reflect the actual system perfectly, as is usually the case in practice. In the literature, several different approaches have been taken to address the impact of such uncertainties on the system performance.

In deterministic control theory, the knowledge is often with respect to modeling errors (such as linearization/parametrization assumptions) and incorrect initializations. Adaptive control and robust control theory for such systems has a rich history ([117, 6, 121, 95]).

A common approach in the literature for linear deterministic systems has been to design controllers that works sufficiently well for all possible uncertain systems under some structured constraints, such as $H_{\infty}$ norm bounded perturbations (see [117, 121]). These techniques provide guaranteed performance for all considered possible models independent of the noise structure and they are deterministic in that the disturbances considered as unknown functions. The design for robust controllers has also been developed through a game theoretic formulation where the minimizer is the controller and the maximizer is the uncertainty ([6]).

In stochastic systems, the knowledge of the system may be with respect to probabilistic aspects, such as the initial prior or the distribution of the noise processes, or due to functional modeling assumptions. All of these considerations can be studied in a unified manner if one develops a probabilistic abstraction and an associated topological analysis on the induced probability measure of a controlled system, where the priors, the transition kernel, and a given policy induce a unique stochastic process (called a strategic measure). The stochastic theory can be used to view the deterministic case as a special setup.
For stochastic systems, the connections between risk-sensitive control and deterministic dynamic games were established in [58, 34]. Using Legendre-type transforms, relative entropy constraints came into the literature to model the uncertainties probabilistically, see e.g. [86, Eqn. (4)] or [34, Eqns. (2)-(3)]. Here, one selects a nominal system which satisfies a relative entropy bound between the actual measure and the nominal measure, solves a risk sensitive optimal control problem, and this solution value provides an upper bound for the original system performance. As such, a common approach in robust stochastic control has been to consider all models which satisfy certain bounds in terms of relative entropy pseudo-distance (or Kullback-Leibler divergence), see e.g. [34, 86, 83] among others. Other metrics or criteria, different from the relative entropy pseudo-distance, have also been used to quantify the uncertainty in the system models. Reference [104] has studied a min-max formulation for robust control where the one-stage transition kernel belongs to a ball under the total variation metric for each state action pair. For distributionally robust stochastic optimization problems, it is assumed that the underlying probability measure of the system lies within an ambiguity set and a worst case single-stage optimization is made considering the probability measures in the ambiguity set. To construct ambiguity sets, [16, 37] use the Wasserstein metric (see Section 2.2), [36] uses the Prokhorov metric which metrizes the weak topology and [69] works with relative entropy.

For fully observed controlled Markov models with finite state spaces, a robust dynamic programming approach has been popular, in particular in the operations research literature. [113, 55, 80] have studied robust dynamic programming approaches through a min-max formulation for fully observed finite state-action space models with uncertain transition probabilities.

Further related work with model uncertainty includes [81, 8], with some further work
1.2. A GENERAL LITERATURE REVIEW AND OUR CONTRIBUTIONS

in the economics literature [49, 46].

Even when the controller has a perfect knowledge of the system, it is often the case that mathematical models designed for control systems are too complex to compute. Therefore, building approximation methods have been a common approach in the literature. Here, we briefly review some approximation results for discrete time stochastic models that are closely related to our setup. For fully observed models, [70] establishes continuity results for approximate models and gives a set convergence result for sets of optimal control actions, however this set convergence result is inconclusive for robustness without further assumptions on the true system model. We also note that approximation methods for stochastic control problems with standard Borel spaces through quantization, which lead to finite models, can be viewed as approximations of transition kernels, but this interpretation requires caution: indeed, [93, 2], among many others, study approximation methods for MDP’s where the convergence of approximate models is satisfied in a particularly constructed fashion. In particular, [93, 92, 94] presents a construction for the approximate models through quantizing the actual model with continuous spaces (leading to a finite space model), which allows for continuity and robustness results with only a weak continuity assumption on the true transition kernel which, in turn, leads to the weak convergence of the approximate models. A detailed analysis of approximation methods for continuous state and action spaces can be found in [90] for both fully observed and partially observed models. However, these positive results on weak convergence of approximate kernels to the true one in such studies do not directly apply to robustness to an arbitrary sequence of models which converges weakly to a true model. In other words, the quantization of state and action spaces results in a special structure for the sequence of approximating transition kernels.
As discussed earlier, stochastic models where the decision maker does not have direct access to the state of the system, and observes the state through a measurement channel, are called partially observed controlled stochastic models, also known as Partially Observed Markov Decision Processes (POMDPs) for Markovian environments. POMDPs provide flexibility for modeling purposes, however, the modeling advantage comes with a mathematically challenging analysis. For POMDPs, working with probability measure valued processes or belief processes via converting the original partially observed stochastic control problem to a fully observed problem, has been a strong analytic tool ([116], [88]). However, computing an optimal policy for this fully observed model, and so for the original POMDP, using classical methods (such as dynamic programming, policy iteration, linear programming) is challenging even if the original system has finite state and action spaces, since the state space of the fully observed model is always uncountable being a space of probability measures.

Due to the complexity of the fully observed reduction of POMDPs, approximation methods are usually needed for optimal control computations. For partially observed setups, [94, 85, 109, 99, 84, 120, 73, 111, 48, 67, 112] are among the works that deal with these approximation techniques for POMDPs.

Similar challenges also remain for learning tasks in POMDPs. If one attempts to learn optimal policies through empirical observations, then the analysis and convergence properties become significantly harder to obtain as the observations progress in a non-Markovian fashion and the belief state space is uncountable. This necessitates finite approximations for POMDPs with rigorous performance guarantees, however, despite the presence of just a few rigorous approximation results, there does not exist a rate of convergence analysis
to our knowledge. This is not surprising since regularity properties for POMDPs are challenging to obtain. Learning in POMDPs is a vast research area and there are numerous studies, however, here we only list [57, 76, 72, 98] as the ones that are most related to the techniques (finite memory approximations) used in this thesis.

A more detailed literature review will be given at every corresponding chapter separately.

Our Contribution

We list our contributions and how they differ from the existing literature as follows:

- **Relaxed metrics for robustness:** We study the robustness problem for discrete time stochastic systems and focus on the uncertainties on the underlying probabilistic model of the system. In particular, we consider the uncertainties on the controlled transition kernel of the system. We use convergence and distance notions for probability measures to quantify the effects of uncertainties. Namely, we analyze the robustness problem using weak convergence, setwise (strong) convergence and convergence under total variation distance. These convergence and distance notions are strictly more relaxed and refined than the distance measures used in existing robust stochastic control literature. In particular, Kullback-Leibler divergence distance, a commonly used distance notion, is a stronger distance notion than the total variation distance through the Pinsker’s inequality ([47, Lemma 5.2.8]) and the Wasserstein distance is a stronger metric than any metric that metrizes the weak convergence topology in general: For compact spaces, the Wasserstein distance of order 1, denoted by $W_1$, metrizes the weak topology (see [108, Theorem 6.9]). For non-compact spaces convergence in the $W_1$ metric implies weak convergence (in particular this metric bounds from above the Bounded-Lipschitz metric [108, p.109]).
• **Explicit convergence rates for POMDP approximations:** For partially observed systems, we establish regularity properties which are usually needed for the analysis and consistency results of POMDPs. For POMDP approximations, we use finite history information variables to construct simpler models by discretizing the belief state space. We provide two discretization schemes for the belief space, both of which have their own advantages. We rigorously establish the near optimality of finite memory feedback control policies for the case where the actions and measurements are finite (with the state being real vector valued), provided that the controlled non-linear filter is stable in a sense to be presented in the relevant chapters of the thesis (Chapter 5). We also explicitly relate the error with the window size. This is the first rigorous result, to our knowledge, where finite window policies are shown to be $\epsilon$-optimal with an explicit rate of convergence with respect to the window size. We emphasize here that our results will apply to all finite state/measurement/action models, though our results are also applicable to setups where the state process is continuous-space valued.

• **Rigorous convergence and near optimality results for finite window Q learning algorithms:** We apply our findings from POMDP approximations to the learning algorithms. In particular, we study the Q-learning problem for POMDPs where one construct the Q iterations using only a finite history of information variables. We establish that these iterations converge to a fixed point equation which coincides with the optimality equations of the approximate models obtained through discretizing the belief models. While there exist many experimental results on Q-learning with finite memory information variables, (i) the rigorous asymptotic convergence (to an approximate MDP value function) for such finite-memory Q-learning algorithms,
1.3. OUTLINE OF THE THESIS

and (ii) the near optimality with an explicit rate of convergence (in the memory size) are results that are new to the literature, to our knowledge.

1.3 Outline of the Thesis

In the first part of the thesis, in Chapter 2 and Chapter 3, we focus on the robustness problem under incorrect system models. In these chapters, the main question we try to answer is the following: Suppose a decision maker designs an optimal policy for an incorrect or incomplete model which he/she thinks is the correct model and applies this policy to the original system. Is it true that the loss occurring from applying the incorrect policy goes to zero, as the incorrect model gets close to the true model? To answer this question, we study the continuity of the optimal cost functions under the convergence of transition kernels, and we analyze the asymptotic behavior of the loss arising from applying the policy designed for the incorrect model. We characterize the convergence of transition models by using notions from convergence of probability measures. Namely, we work with weak convergence, setwise convergence and total variation convergence.

In Chapter 2, we focus on the infinite horizon discounted cost problems. We analyze both fully observed and partially observed setups. We show that the continuity of the optimal cost can be established under total variation convergence of the transition kernels under mild assumptions and with further restrictions on the dynamics and observation model under weak and setwise convergence of the transition kernels. Using these continuity properties, we establish convergence results and error bounds due to mismatch that occurs by the application of a control policy which is designed for an incorrectly estimated system model to a true model, thus establishing positive and negative results on robustness.

In Section 2.7, we view finite model approximations for fully observed models with
continuous state spaces as a particular instance of robustness problem. We show how in particular, continuous weak convergence applies under finite model quantization of weakly continuous kernels and thus, the findings from Section 2.6 can be utilized to arrive at consistency results for finite approximations.

In Section 2.8 we consider cases where the system models are constructed empirically. Let $T(\cdot|x,u)$ be a transition kernel given previous state and action variables $x \in X$, $u \in U$, which is unknown to the decision maker (DM). Suppose the DM builds a model for the transition kernels, $T_n(\cdot|x,u)$, for all possible $x \in X$, $u \in U$ by collecting training data (e.g. from the evolving system). Do we have that the cost calculated under $T_n$ converges to the true cost (i.e., do we have that the cost obtained from applying the optimal policy for the empirical model converges to the true cost as the training length increases)? We provide different cases for which the empirically constructed models satisfy certain convergence properties presented throughout Chapter 2.

In Chapter 3 we study the robustness problem under the infinite horizon average cost setup. We show that for fully observed models, sufficient conditions presented in Chapter 2 for discounted cost problems are in general not sufficient to ensure robustness for average cost problems. However, we show that the average optimal cost is continuous in the convergence of controlled transition kernel models where convergence of models entails (i) continuous weak convergence in state and actions, and (ii) continuous setwise convergence in the actions for every fixed state variable, in addition to either uniform ergodicity or some regularity conditions. We establish that the mismatch error due to the application of a control policy designed for an incorrectly estimated model to the true model decreases to zero as the incorrect model approaches the true model under the stated convergence criteria. Furthermore, in Section 3.7 we apply our findings to the models learned using empirical
In the second part of the thesis, we focus on partially observed Markov decision processes (POMDPs). We first analyze regularity properties of POMDPs (Chapter 4), we then provide approximation techniques for partially observed models (Chapter 5) and we finally study approximate learning algorithms for POMDPs (Chapter 6).

In Chapter 4, we first establish some topological properties for partially observed models which are useful for consistency results for approximation techniques. A well known result states that any partially observed controlled model can be reduced to a fully observed controlled model whose state process is probability measure valued ([116], [88]). In parts of the thesis, we will observe that an important condition for the fully observed models, is the weak continuity of their transition kernels. In this chapter, we find sufficient conditions on the partially observed model which make the transition kernel of its reduced model weakly continuous. In particular, we prove that the transition kernel of the belief MDP is weakly continuous (i) first under weak continuity of the transition probability of controlled partially observed Markov chain and total variation continuity of its observation channel, and then, (ii) under total variation continuity of the transition probability of controlled partially observed Markov chain without any restrictions on the measurement model.

In Chapter 5, we present two approximation methods for POMDPs. As we have stated before, POMDPs can be converted to fully observed controlled problems with probability measure valued state processes where the new state process is called the belief state. This reduced model, in general, is analytically and computationally demanding. Hence, approximation techniques are needed for POMDPs. In Chapter 5, we focus on two techniques that use finite window information variables to construct finite approximations for the belief process by discretizing the corresponding belief spaces. By finite window information
variables, we refer to a fixed length of history variables observed by the decision maker. Both methods rely on a finite subset of the belief state space that consists of Bayesian updates of a fixed initial distribution under the finite window information variables. We present the first method in Section 5.3 which makes use of a nearest neighbor map to choose the closest element to the original belief state from the finite subset to construct the approximate model. In this section, we establish error bounds that vanish exponentially fast with increasing window size under some filter stability conditions. The second method is presented in Section 5.4 in which, instead of mapping the original belief state to the closest element in the finite subset, we map the original state to the element with matching finite history. We provide convergence results for the approximation error with increasing window size for this technique as well. The second method results in coarser error bounds compared to the approximation presented in Section 5.3 however, it provides a considerable practical advantage that will be evident in Chapter 6. In Section 5.5 we present a detailed comparison of the methods presented in Section 5.3 and 5.4. In Section 5.6 we provide numerical examples and simulation results that verify the proven results in this chapter.

In Chapter 6 we focus on approximate Q learning algorithms for partially observed controlled models and for fully observed models with continuous state spaces. Traditional Q learning algorithms in the literature are proven to converge to optimal fixed point equations for models with finite state spaces whose states can be observed perfectly. However, for partially observed models, the convergence, in general, is not guaranteed since the observation process is no longer controlled Markov, a property that is heavily used for convergence results. Even when convergence occurs, what the limit operationally means is also a question to be answered. For continuous state spaces, the usual Q learning algorithms are
again not directly applicable, since it is not feasible to track uncountably many Q values.

In Section 6.4 for POMDPs, we consider the convergence of a Q learning algorithm for control policies using a finite history of past observations and control actions, and, consequentially, we establish near optimality of such limit Q functions under explicit filter stability conditions. We use explicit error bounds relating the approximation error to the length of the finite history window from Section 5.4. We establish the convergence of such Q-learning iterations under mild ergodicity assumptions on the state process during the exploration phase. We further show that the limit fixed point equation gives an optimal solution for an approximate belief-MDP constructed in Section 5.4. We then provide bounds on the performance of the policy obtained using the limit Q values compared to the performance of the optimal policy for the POMDP that could be calculated if the system was fully known, where we also present explicit conditions using recent results on filter stability in controlled POMDPs. While there exist many experimental results, (i) the rigorous asymptotic convergence (to an approximate MDP value function) for such finite-memory Q-learning algorithms, and (ii) the near optimality with an explicit rate of convergence (in the memory size) are results that are new to the literature, to our knowledge.

In Section 6.6 we study the Q learning problem for fully observed models with continuous state spaces. For MDPs with continuous state spaces, it is not feasible to follow the usual Q learning algorithm as there are uncountably many Q values to approximate. We show that the problem can be seen as a POMDP with a quantizer channel if we discretize the state space and thus, we can use the results from Section 6.4 to build an approximate Q learning algorithm.

Finally, in Chapter 7 we conclude the thesis by giving a summary of the results and providing possible research directions.
1.3. OUTLINE OF THE THESIS

Some of the results can be found as stand alone papers in [61, 64, 63, 65, 60].

We also note that, in [62], we have studied robustness and continuity properties for partially observed stochastic control problems with respect to the prior measures with fixed transition kernels. However, due to space constraints we chose not to include it in the thesis.
Chapter 2

Continuity and Robustness under Incorrect Transition Kernels under Infinite Horizon Discounted Cost Criterion

In this chapter, we study continuity properties of discrete-time stochastic control problems with respect to system models (i.e., controlled transition kernels) and robustness of optimal control policies designed for incorrect models applied to the true system. We focus on the infinite horizon discounted expected cost criterion for both fully observed and partially observed setups.

We first restate the definitions for the problems we will be studying in this chapter. The discounted infinite horizon cost function is defined as:

\[ J_\beta(P, T, \gamma) = E^T_P \sum_{t=0}^{\infty} \beta^t c(X_t, U_t), \quad \beta \in (0, 1) \]

over the set of admissible policies \( \gamma \in \Gamma \), where \( c : X \times U \to \mathbb{R} \) is a Borel-measurable stage-wise cost function and \( E^T_P, \gamma \) denotes the expectation with initial state probability measure \( P \) and transition kernel \( T \) under policy \( \gamma \).
We define the optimal cost as a function of the priors and the transition kernels as

\[ J^*_\beta(P, \mathcal{T}) = \inf_{\gamma \in \Gamma} J_\beta(P, \mathcal{T}, \gamma). \]

The focus will be to address the following problems:

**Problem P1: Continuity of \( J^*_\beta(P, \mathcal{T}) \) under the convergence of the transition kernels.**

Let \( \{\mathcal{T}_n, n \in \mathbb{N}\} \) be a sequence of transition kernels which converge in some sense to another transition kernel \( \mathcal{T} \). Does this imply that

\[ J^*_\beta(P, \mathcal{T}_n) \to J^*_\beta(P, \mathcal{T}), \]

**Problem P2: Robustness to incorrect models.** A problem of major practical importance is robustness of an optimal controller to modeling errors. Suppose that an optimal policy is constructed according to a model which is incorrect: how does the application of the control to the true model affect the system performance and does the error decrease to zero as the models become closer to each other? In particular, suppose that \( \gamma^*_n \) is an optimal policy designed for \( \mathcal{T}_n \), an incorrect model for a true model \( \mathcal{T} \). Is it the case that if \( \mathcal{T}_n \to \mathcal{T} \) then \( J_\beta(P, \mathcal{T}, \gamma^*_n) \to J^*_\beta(P, \mathcal{T})? \)

2.1 Literature Review

Robustness is a desired property for the optimal control of stochastic or deterministic systems when a given model does not reflect the actual system perfectly, as is usually the case in practice.

A common approach in the literature has been to design controllers that work sufficiently well for all possible uncertain systems under some structured constraints, such as
2.1. LITERATURE REVIEW

$H_\infty$ norm bounded perturbations (see \cite{6,121}). The design for robust controllers has often been developed through a game theoretic formulation where the minimizer is the controller and the maximizer is the uncertainty. The connections of this formulation to risk sensitive control were established in \cite{58,34}. Using Legendre-type transforms, relative entropy constraints came in to the literature to probabilistically model the uncertainties, see e.g. \cite{86,Eqn. (4)} or \cite{34,Eqns. (2)-(3)]. Here, one selects a nominal system which satisfies a relative entropy bound between the actual measure and the nominal measure, solves a risk sensitive optimal control problem, and this solution value provides an upper bound for the original system performance. As such, a common approach in robust stochastic control has been to consider all models which satisfy certain bounds in terms of relative entropy pseudo-distance (or Kullback-Leibler divergence), see e.g. \cite{34,86,83} among others.

Other metrics or criteria, different from the relative entropy pseudo-distance, have also been used to quantify the uncertainty in the system models. Reference \cite{104} has studied a min-max formulation for robust control where the one-stage transition kernel belongs to a ball under the total variation metric for each state action pair. For distributionally robust stochastic optimization problems, it is assumed that the underlying probability measure of the system lies within an ambiguity set and a worst case single-stage optimization is made considering the probability measures in the ambiguity set. To construct ambiguity sets, \cite{16,37} use the Wasserstein metric (see Section 2.2), \cite{36} uses the Prokhorov metric which metrizes the weak topology and \cite{69} works with relative entropy. \cite{113,55,80} have studied robust dynamic programming approaches through a min-max formulation for fully observed finite state-action space models with uncertain transition probabilities. Further related work with model uncertainty includes \cite{81,8}, with some further work in the economics literature \cite{49,46}.
For fully observed models, [70, Theorem 5.1] establishes continuity results for approximate models and gives a set convergence result for sets of optimal control actions, however this set convergence result is inconclusive for robustness without further assumptions on the true system model. Reference [79] is another related work which studies continuity of the value function for fully observed models under a general metric defined as the integral probability metric which captures both the total variation metric or the Kantorovich metric with different setups (which is not weaker than the metrics leading to weak convergence). We also note that approximation methods for stochastic control problems with standard Borel spaces through quantization, which leads to finite models, can be viewed as approximations of transition kernels, but this interpretation requires caution: indeed, [93, 2], among many others, study approximation methods for MDP’s where the convergence of approximate models is satisfied in a particularly constructed fashion. In particular, [93] presents a construction for the approximate models through quantizing the actual model with continuous spaces (leading to a finite space model), which allows for continuity and robustness results with only a weak continuity assumption on the true transition kernel which, in turn, leads to the weak convergence of the approximate models. A detailed analysis of approximation methods for continuous state and action spaces can be found in [90] for both fully observed and partially observed models. However, these positive results on weak convergence of approximate kernels to the true one in such studies do not directly apply to robustness to an arbitrary sequence of models which converges weakly to a true model, as our counterexamples in this chapter will demonstrate.

In our analysis, we will study various convergence criteria. Our findings involving weak convergence are related to the following convergence notion: Let a stochastic process converge to another one in the following sense: all finite dimensional marginals converge...
weakly and the conditional kernels on the future random variables given the past converge weakly as well. This has been termed as *extended weak convergence* [1] or convergence under the *information topology* [50]; these have recently shown to be equivalent in discrete-time [4, Theorem 1.1]. Applications of this notion of convergence to robustness to system models in stochastic control have recently been studied in different contexts [7, 5]. In our case, we will relate this to a generalized (dominated) convergence criterion under *continuous weak convergence* [70, Theorem 3.5] and [97, Theorem 3.5].

In Section 2.3 we will present the contributions, which will also make the comparison with the reviewed literature more explicit.

### 2.2 Some Examples and Convergence Criteria for Transition Kernels

#### 2.2.1 Convergence Criteria for Transition Kernels

Before presenting convergence criteria for controlled transition kernels, we first review convergence of probability measures. Three important notions of convergences for sets of probability measures to be studied are weak convergence, setwise convergence and convergence under total variation. For $N \in \mathbb{N}$, a sequence $\{\mu_n, n \in \mathbb{N}\}$ in $\mathcal{P}(\mathbb{R}^N)$ is said to converge to $\mu \in \mathcal{P}(\mathbb{R}^N)$ *weakly* if

$$\int_{\mathbb{R}^N} c(x) \mu_n(dx) \rightarrow \int_{\mathbb{R}^N} c(x) \mu(dx) \tag{*}$$

for every continuous and bounded $c : \mathbb{R}^N \rightarrow \mathbb{R}$. $\{\mu_n\}$ is said to converge *setwise* to $\mu \in \mathcal{P}(\mathbb{R}^N)$ if (*) holds for all measurable and bounded $c : \mathbb{R}^N \rightarrow \mathbb{R}$. For probability
measures \( \mu, \nu \in \mathcal{P}(\mathbb{R}^N) \), the total variation metric is given by

\[
\|\mu - \nu\|_{TV} = 2 \sup_{B \in \mathcal{B}(\mathbb{R}^N)} |\mu(B) - \nu(B)| = \sup_{f : \|f\|_{\infty} \leq 1} \left| \int f(x) \nu(dx) - \int f(x) \mu(dx) \right|
\]

where the supremum is taken over all measurable real \( f \) such that \( \|f\|_{\infty} = \sup_{x \in \mathbb{R}^N} |f(x)| \leq 1 \). A sequence \( \{\mu_n\} \) is said to converge in total variation to \( \mu \in \mathcal{P}(\mathbb{R}^N) \) if \( \|\mu_n - \mu\|_{TV} \to 0 \).

Total variation defines a stringent metric for convergence; for example, a sequence of discrete probability measures does not converge in total variation to a probability measure which admits a density function. Setwise convergence, though, induces a topology on the space of probability measures which is not metrizable. However, the space of probability measures on a complete, separable, metric (Polish) space endowed with the topology of weak convergence is itself complete, separable and metric [82]. We also note here that relative entropy convergence, through Pinsker’s inequality [47, Lemma 5.2.8], is stronger than even total variation convergence which has also been studied in robust stochastic control as reviewed earlier. Another metric for probability measures is the Wasserstein distance: For compact spaces, the Wasserstein distance of order 1, denoted by \( W_1 \), metrizes the weak topology (see [108, Theorem 6.9]). For non-compact spaces convergence in the \( W_1 \) metric implies weak convergence (in particular this metric bounds from above the Bounded-Lipschitz metric [108, p.109]). Furthermore, instead of using the usual metric, say \( d \), on \( X \), one can always construct a bounded metric, e.g. \( \tilde{d} = d/(1 + d) \), for which the induced Wasserstein distance metrizes the weak convergence topology on \( \mathcal{P}(X) \) [108, p.109, Corollary 6.13]. Considering these relations, our results in this section can be directly generalized to the relative entropy distance or the Wasserstein distance. Building on the above, we introduce the following convergence notions for (controlled) transition kernels.

**Definition 2.2.1.** For a sequence of transition kernels \( \{\mathcal{T}_n, n \in \mathbb{N}\} \), we say that
2.2. SOME EXAMPLES AND CONVERGENCE CRITERIA FOR TRANSITION KERNELS

(i) $T_n \to T$ weakly if $T_n(\cdot|x,u) \to T(\cdot|x,u)$ weakly, for all $x \in X$ and $u \in U$.

(ii) $T_n \to T$ setwise if $T_n(\cdot|x,u) \to T(\cdot|x,u)$ setwise, for all $x \in X$ and $u \in U$.

(iii) $T_n \to T$ under the total variation distance if $T_n(\cdot|x,u) \to T(\cdot|x,u)$ under total variation, for all $x \in X$ and $u \in U$.

2.2.2 Examples

Let a controlled model be given as

$$x_{t+1} = F(x_t, u_t, w_t),$$

where $\{w_t\}$ is an i.i.d. noise process. The uncertainty on the transition kernel for such a system may arise from lack of information on $F$ or the i.i.d. noise process $w_t$ or both:

(i) Let $\{F_n\}$ denote an approximating sequence for $F$, so that $F_n(x, u, w) \to F(x, u, w)$ pointwise. Assume that the probability measure of the noise is known. Then, corresponding kernels $T_n$ converges weakly to $T$: If we denote the probability measure of $w$ with $\mu$, for any $g \in C_b(X)$ and for any $(x_0, u_0) \in X \times U$ using the dominated convergence theorem we have

$$\lim_{n \to \infty} \int g(x_1)T_n(dx_1|x_0, u_0) = \lim_{n \to \infty} \int g(F_n(x_0, u_0, w))\mu(dw)$$

$$= \int g(F(x_0, u_0, w))\mu(dw) = \int g(x_1)T(dx_1|x_0, u_0).$$

(ii) Much of the robust control literature deals with deterministic systems where the nominal model is a deterministic perturbation of the actual model (see e.g. [95]). The considered model is in the following form; $\tilde{F}(x_t, u_t) = F(x_t, u_t) + \Delta F(x_t, u_t),$. 
where $F$ represents the nominal model and $\Delta F$ is the model uncertainty satisfying some norm bounds. For such deterministic systems, pointwise convergence of $\tilde{F}$ to the nominal model $F$, i.e. $\Delta F(x_t, u_t) \to 0$, can be viewed as weak convergence for deterministic systems by the discussion in (i). It is evident, however, that total variation convergence would be too strong for such a convergence criterion, since $\|\delta_{\tilde{F}(x_t, u_t)} - \delta_{F(x_t, u_t)}\|_{TV} = 2$ for all $\Delta F(x_t, u_t) \neq 0$.

(iii) Let $F(x_t, u_t, w_t) = f(x_t, u_t) + w_t$ be such that the function $f$ is known and $w_t \sim \mu$ is not known correctly and an incorrect model $\mu_n$ is assumed. If $\mu_n \to \mu$ weakly, setwise, or total variation, then the corresponding transition kernels $T_n$ converges in the same sense to $T$. Observe the following,

$$
\int g(x_1) T_n(dx_1|x_0, u_0) - \int g(x_1) T(dx_1|x_0, u_0) = \int g(w_0 + f(x_0, u_0))\mu_n(dw_0) - \int g(w_0 + f(x_0, u_0))\mu(dw_0). \tag{2.1}
$$

a) Suppose $\mu_n \to \mu$ weakly. If $g$ is a continuous and bounded function then $g(\cdot + f(x_0, u_0))$ is a continuous and bounded function for all $(x_0, u_0) \in X \times U$. Thus, (2.1) goes to 0. Note that $f$ does not need to be continuous. b) Suppose $\mu_n \to \mu$ setwise. If $g$ is a measurable and bounded function, then $g(\cdot + f(x_0, u_0))$ measurable and bounded for all $(x_0, u_0) \in X \times U$. Thus, (2.1) goes to 0. c) Finally, assume $\mu_n \to \mu$ in total variation. If $g$ is bounded, (2.1) converges to 0, as in item (b). As a special case, assume that $\mu_n$ and $\mu$ admit densities $h_n$ and $h$ respectively; then the pointwise convergence of $h_n$ to $h$ implies the convergence of $\mu_n$ to $\mu$ in total variation by Scheffé’s Theorem.

(iv) Suppose now neither $F$ nor the probability model of $w_t$ is known perfectly. It is
assumed that \( w_t \) admits a measure \( \mu_n \) and \( \mu_n \rightarrow \mu \) weakly. For the function \( F \) we again have an approximating sequence \( \{ F_n \} \). If \( F_n(x, u, w_n) \rightarrow F(x, u, w) \) for all \( (x, u) \in X \times U \) and for any \( w_n \rightarrow w \), then the transition kernel \( T_n \) corresponding to the model \( F_n \) converges weakly to the one of \( F, T \): For any \( g \in C_b(X) \),

\[
\lim_{n \to \infty} \int g(x_1) T_n(dx_1|x_0, u_0) = \lim_{n \to \infty} \int g(F_n(x_0, u_0, w)) \mu_n(dw) = \int g(F(x_0, u_0, w)) \mu(dw) = \int g(x_1) T(dx_1|x_0, u_0).
\]

In the analysis above we used a generalized dominated convergence result, Lemma 2.9.1, to be presented later building on [70, Theorem 3.5] and [97, Theorem 3.5].

(v) Let again \( \{ F_n \} \) denote an approximating sequence for \( F \) and suppose now \( F_{x_0, u_0, n}(\cdot) := F_n(x_0, u_0, \cdot) : W \to X \) is invertible for all \( x_0, u_0 \in X \times U \) and \( F^{-1}_{(x_0, u_0), n}(\cdot) \) denotes the inverse for fixed \( (x_0, u_0) \). It is assumed that \( F^{-1}_{(x_0, u_0), n}(x_1) \rightarrow F^{-1}_{x_0, u_0}(x_1) \) pointwise for all \( (x_0, u_0) \). Suppose further that the noise process \( w_t \) admits a continuous density \( f_W(w) \). The Jacobian matrix, \( \frac{\partial x_1}{\partial w} \), is the matrix whose components are the partial derivatives of \( x_1 \), i.e. with \( x_1 \in X \subset \mathbb{R}^m \) and \( w \in W \subset \mathbb{R}^m \), it is an \( m \times m \) matrix with components \( \frac{\partial(x_1)_i}{\partial w_j}, 1 \leq i, j \leq m \). If the Jacobian matrix of derivatives \( \frac{\partial x_1}{\partial w}(w) \) is continuous in \( w \) and nonsingular for all \( w \) then by the inverse function theorem of vector calculus, we have that the density of the state variables can be written as

\[
f_{X_{x_0, u_0}}(x_1) = f_W(F^{-1}_{x_0, u_0, n}(x_1)) \left| \frac{\partial x_1}{\partial w}(F^{-1}_{x_0, u_0, n}(x_1)) \right|^{-1}
\]

\[
f_{X_{x_0, u_0}}(x_1) = f_W(F^{-1}_{x_0, u_0}(x_1)) \left| \frac{\partial x_1}{\partial w}(F^{-1}_{x_0, u_0}(x_1)) \right|^{-1}.
\]
With the above, \( f_{X_1,n(x_0,u_0)}(x_1) \to f_{X_1(x_0,u_0)}(x_1) \) pointwise for all fixed \((x_0,u_0)\).

Therefore, by Scheffé’s Theorem, the corresponding kernels \( T_n(\cdot|x_0,u_0) \to T(\cdot|x_0,u_0) \)
in total variation for all \((x_0,u_0)\).

### 2.3 Summary of Results and Contributions of the Chapter

We now introduce the main assumptions that will be occasionally used for our technical results in this section.

**Assumption 2.3.1.**

(a) The sequence of transition kernels \( T_n \) satisfies the following:

\[ \{T_n(\cdot|x_n,u_n), n \in \mathbb{N}\} \text{ converges weakly to } T(\cdot|x,u) \] for any sequence \( \{x_n,u_n\} \subset X \times U \) and \( x,u \in X \times U \) such that \( (x_n,u_n) \to (x,u) \).

(b) The stochastic kernel \( T(\cdot|x,u) \) is weakly continuous in \((x,u)\).

(c) The stage-wise cost function \( c(x,u) \) is non-negative, bounded and continuous on \( X \times U \).

(d) \( U \) is compact.

**Assumption 2.3.2.**

- The observation channel \( Q(\cdot|x) \) is continuous in total variation i.e., if \( x_k \to x \), then \( Q(\cdot|x_k) \to Q(\cdot|x) \) in total variation (only for partially observed models).

**Assumption 2.3.3.**

(a) The sequence of transition kernels \( T_n \) satisfies the following:

\[ \{T_n(\cdot|x,u_n), n \in \mathbb{N}\} \text{ converges setwise to } T(\cdot|x,u) \] for any sequence \( \{u_n\} \subset U \) and \( x,u \in X \times U \) such that \( u_n \to u \).

(b) The stochastic kernel \( T(\cdot|x,u) \) is setwise continuous in \( u \),
2.3. SUMMARY OF RESULTS AND CONTRIBUTIONS OF THE CHAPTER

(c) The stage-wise cost function \( c(x, u) \) is non-negative, bounded and continuous on \( U \).

(d) \( U \) is compact.

**Assumption 2.3.4.**  (a) The sequence of transition kernels \( T_n \) satisfies the following:
\[
\|T_n(\cdot|x, u_n) - T(\cdot|x, u)\|_{TV} \to 0 \quad \text{for any sequence } \{u_n\} \subset U \text{ and } x, u \in X \times U
\]
such that \( u_n \to u \).

(b) The stochastic kernel \( T(\cdot|x, u) \) is continuous in total variation in \( u \).

(c) The stage-wise cost function \( c(x, u) \) is non-negative, bounded and continuous on \( U \).

(d) \( U \) is compact.

**Remark 2.3.1.** The following examples provide some sufficient conditions for the above technical assumptions:

- Suppose the channel and the state dynamics have the form: \( y_t = g(x_t) + v_t \), and \( x_{t+1} = f(x_t, u_t) + w_t \). If \( g \) and \( f \) are continuous and if \( v_t \) and \( w_t \) admit continuous probability density functions, then the kernel \( \mathcal{Q}(dy|x) \) induced by the channel and the kernel \( T(dx_1|x_0, u_0) \) induced by the dynamics are continuous in total variation by an application of Scheffe’s Lemma [13, Theorem 16.12].

- Suppose the dynamics are given by \( x_{t+1} = F(x_t, u_t, w_t) \) where \( F(x, u, w) \) is continuous in \( (x, u) \), then the corresponding kernel \( T(dx_1|x_0, u_0) \) is weakly continuous in \( (x_0, u_0) \) regardless of the distribution of \( w_t \).

The reader can refer to Section 4.4.2 for a more detailed discussion on the assumptions and the examples presented in this section.
In Section 2.4 and 2.5 we study continuity (Problem P1) and robustness (Problem P2) for partially observed models. In particular we show that

- Continuity and robustness do not hold in general under weak convergence of kernels (Theorem 2.4.1).

- Under Assumptions 2.3.1 and 2.3.2, continuity and robustness hold (Theorem 2.4.3, Theorem 2.5.3).

- Continuity and robustness do not hold in general under setwise convergence of the kernels (Theorem 2.4.4).

- Continuity and robustness do not hold in general under total variation convergence of the kernels (Example 2.6.1).

- Under Assumption 2.3.4, continuity and robustness hold (Theorem 2.4.5, Theorem 2.5.2).

In Section 2.6, we study continuity (Problem P1) and robustness (Problem P2) for fully observed models. In particular we show that

- Continuity and robustness do not hold in general under weak convergence of kernels (Theorem 2.6.1, Example 2.6.1).

- Under Assumption 2.3.1, continuity holds (Theorem 2.6.2), under Assumption 2.3.1, robustness holds if the optimal policies for every initial point are identical (Theorem 2.6.4).

- Continuity and robustness do not hold in general under setwise convergence of the kernels (Theorem 2.6.5, Theorem 2.6.7).
2.4. CONTINUITY OF OPTIMAL COST WITH RESPECT TO CONVERGENCE OF TRANSITION KERNELS (PARTIALLY OBSERVED CASE)

- Under Assumption 2.3.3, continuity holds (Theorem 2.6.6), under Assumption 2.3.3, robustness holds if the optimal policies for every initial point are identical (Theorem 2.6.8).

- Continuity and robustness do not hold in general under total variation convergence of the kernels (Example 2.6.1).

- Under Assumption 2.3.4, continuity and robustness hold (Subsection 2.6.3).

Compared to the existing literature reviewed earlier, the above results use strictly more relaxed and refined convergence criteria to study robustness. In Section 2.8, these results will be applied to arrive at positive and negative implications on empirical consistency and data-driven learning in stochastic control since often system models are learned through empirical training data where typically weak convergence criterion applies (in an almost sure sense) but stronger convergence criteria do not.

2.4 Continuity of Optimal Cost with respect to Convergence of Transition Kernels (Partially Observed Case)

In this section, we will study continuity of the optimal discounted cost under the convergence of transition kernels for partially observed models.

2.4.1 Weak Convergence

Absence of Continuity under Weak Convergence

The following shows that the optimal cost may not be continuous under weak convergence of transition kernels.
Theorem 2.4.1. Let $\mathcal{T}_n \to \mathcal{T}$ weakly, then it is not necessarily true that $J^*_\beta(P, \mathcal{T}_n) \to J^*_\beta(P, \mathcal{T})$ even when the prior distributions are same, the measurement channel $Q$ is continuous in total variation and $c(x, u)$ is continuous and bounded on $X \times U$.

Proof. We prove the result with a counterexample. Let $X = U = Y = [-1, 1]$ and $c(x, u) = (x - u)^2$, the observation channel is chosen to be uniformly distributed over $[-1, 1]$, $Q \sim U([-1, 1])$, the initial distributions of the state variable are chosen to be same as $P \sim \delta_1$ where $\delta_A(x) := 1_{\{x \in A\}}$ for Borel $A$, and the transition kernels are:

$$
\mathcal{T}(\cdot|x, u) = \delta_{-1}(x)[1/2 \delta_1(\cdot) + 1/2 \delta_{-1}(\cdot)] + \delta_1(x)[1/2 \delta_1(\cdot) + 1/2 \delta_{-1}(\cdot)] + (1 - \delta_{-1}(x))(1 - \delta_1(x))\delta_0(\cdot)
$$

$$
\mathcal{T}_n(\cdot|x, u) = \delta_{-1}(x)[1/2 \delta_{1-1/n}(\cdot) + 1/2 \delta_{-1+1/n}(\cdot)] + \delta_1(x)[1/2 \delta_{1-1/n}(\cdot) + 1/2 \delta_{-1+1/n}(\cdot)] + (1 - \delta_{-1}(x))(1 - \delta_1(x))\delta_0(\cdot).
$$

In other words, for $\mathcal{T}$:

If $x_t \in \{1, -1\}$ then $x_{t+1} = 1$ or $-1$

else $x_{t+1} = 0$

for $\mathcal{T}_n$

If $x_t \in \{1, -1\}$ then $x_{t+1} = 1 - \frac{1}{n}$ or $-1 + \frac{1}{n}$

else $x_{t+1} = 0$

independent of the control, where the events noted above with or are equally likely. It can be seen that $\mathcal{T}_n \to \mathcal{T}$ weakly according to Definition 2.2.1(i). Since the cost function is mean square error, control does not affect the dynamics and the channel is non-informative,
the optimal policy is:
\[
\gamma_k^*(y_{[0,k]}) = E[X_k] = \begin{cases} 
0 & \text{if } k > 0 \\
1 & \text{if } k = 0.
\end{cases}
\]

Note that the cost function is continuous, and the measurement channel is continuous in total variation. The optimal discounted costs can be found as:

\[
J_{\beta}^*(P, \mathcal{T}) = \sum_{k=1}^{\infty} E_P^T[\beta^k X_k^2] = \sum_{k=1}^{\infty} \beta^k = \frac{\beta}{1 - \beta}
\]

\[
J_{\beta}^*(P, \mathcal{T}_n) = \sum_{k=1}^{\infty} E_P^T[\beta^k X_k^2] = \beta \left[ \frac{1}{2} (1 - \frac{1}{n})^2 + \frac{1}{2} (-1 + \frac{1}{n})^2 \right].
\]

Then we have \( J_{\beta}^*(P, \mathcal{T}_n) \to \beta \neq \frac{\beta}{1 - \beta} \)

2.4.2 A Sufficient Condition for Continuity under Weak Convergence

In the following, we will establish and utilize some regularity properties for the optimal cost with respect to the convergence of transition kernels.

**Assumption 2.4.1.** (a) The stochastic kernel \( \mathcal{T}(\cdot|x, u) \) is weakly continuous in \((x, u)\).

(b) The observation channel \( Q(\cdot|x) \) is continuous in total variation i.e., if \( x_k \to x \), then \( Q(\cdot|x_k) \to Q(\cdot|x) \) in total variation.

(c) The stage-wise cost function \( c(x, u) \) is non-negative, bounded and continuous on \( X \times U \)

(d) \( U \) is compact.
It is a standard result that any partially observed Markov decision process (POMDP) can be reduced to a (completely observable) MDP, whose states are the posterior state distributions or beliefs of the observer; that is, the state at time $t$ is $Z_t(\cdot) := \Pr\{X_t \in \cdot | Y_0, \ldots, Y_t, U_0, \ldots, U_{t-1}\} \in \mathcal{P}(X)$. We call this equivalent MDP the belief-MDP. The belief-MDP has state space $Z = \mathcal{P}(X)$ and action space $U$. Under the topology of weak convergence, since $X$ is a Borel space, $Z$ is metrizable with the Prokhorov metric which makes $Z$ into a Borel space [82]. The transition probability $\eta$ of the belief-MDP can be constructed through non-linear filtering equations. The one-stage cost function $c$ of the belief-MDP is given by $\tilde{c}(z, u) := \int_X c(x, u) z(dx)$. By [10] Proposition 7.30, the one stage cost function $\tilde{c}$ of the belief-MDP is continuous and bounded, that is in $C_b(Z \times U)$, under Assumption 2.4.1-(c). By [42] Theorem 3.7, Example 4.1] (see also Chapter 4), under Assumption 2.4.1, the stochastic kernel $\eta$ for belief-MDP is weakly continuous in $(z, u)$. For an MDP with weakly continuous transition probabilities and compact action spaces, it follows that an optimal control policy exists: This follows because the discounted cost optimality operator $T : C_b(Z) \rightarrow C_b(Z)$ (see e.g. [53] Chapter 8.5]):

$$
(T(f))(z) = \min_{u} (\tilde{c}(z, u) + \beta E[f(z_1) | z_0 = z, u_0 = u])
$$

(2.2)

is a contraction from $C_b(Z)$ to itself under the supremum norm. As a result, there exists a fixed point, the value function, and an optimal control policy exists. In view of this existence result, in the following we will consider optimal policies. We note though that for the results which do not use the assumption, one may use $\epsilon$-optimal policies without affecting the results.

Theorem 2.4.2. Under Assumptions [2.3.1] and [2.3.2] that is
2.4. CONTINUITY OF OPTIMAL COST WITH RESPECT TO CONVERGENCE OF TRANSITION KERNELS (PARTIALLY OBSERVED CASE)

• The sequence of transition kernels $T_n$ satisfies the following: $\{T_n(\cdot|x_n, u_n), n \in \mathbb{N}\}$ converges weakly to $T(\cdot|x, u)$ for any sequence $\{x_n, u_n\} \subset X \times U$ and $x, u \in X \times U$ such that $(x_n, u_n) \to (x, u)$.

• The stochastic kernel $T(\cdot|x, u)$ is weakly continuous in $(x, u)$.

• The observation channel $Q(\cdot|x)$ is continuous in total variation i.e., if $x_k \to x$, then $Q(\cdot|x_k) \to Q(\cdot|x)$ in total variation (only for partially observed models).

• The stage-wise cost function $c(x, u)$ is non-negative, bounded and continuous on $X \times U$.

• $U$ is compact.

\[
\sup_{\gamma \in \Gamma} |J_\beta(P, T_n, \gamma) - J_\beta(P, T, \gamma)| \to 0.
\]

Proof.

\[
\sup_{\gamma \in \Gamma} |J_\beta(P, T_n, \gamma) - J_\beta(P, T, \gamma)| = \sup_{\gamma \in \Gamma} \left| \sum_{t=0}^{\infty} \beta^t \left( E_P^T \left[ c(X_t, \gamma(Y_{[0,t]})) \right] - E_P^{T_n} \left[ c(X_t, \gamma(Y_{[0,t]})) \right] \right) \right| \\
\leq \sum_{t=0}^{\infty} \beta^t \sup_{\gamma \in \Gamma} \left| E_P^T \left[ c(X_t, \gamma(Y_{[0,t]})) \right] - E_P^{T_n} \left[ c(X_t, \gamma(Y_{[0,t]})) \right] \right|.
\]

Recall that an admissible policy $\gamma$ is a sequence of control functions $\{\gamma_t, t \in \mathbb{Z}_+\}$. At the last step above, we make a slight abuse of notation; the $\sup$ at the first step is over all sequence of control functions $\{\gamma_t, t \in \mathbb{Z}_+\}$ whereas the $\sup$ at the last step is over all sequence of control functions $\{\gamma_{t'}, t' \leq t\}$ but we will use the same notation, $\gamma$, in the rest of the proof. In Appendix 2.9.1, we show the following for any $t \geq 0$:

\[
\sup_{\gamma \in \Gamma} \left| E_P^T \left[ c(X_t, \gamma(Y_{[0,t]})) \right] - E_P^{T_n} \left[ c(X_t, \gamma(Y_{[0,t]})) \right] \right| \to 0.
\] (2.3)
For any $\epsilon > 0$, we choose a $K < \infty$ such that $\sum_{t=K+1}^{\infty} \beta^k 2\|c\|_\infty \leq \epsilon/2$. For the chosen $K$, we choose an $N < \infty$ such that

$$\sup_{\gamma \in \Gamma} \left| E_P^T \left[ c(X_t, \gamma(Y_{[0,t]})) \right] - E_P^{T_n} \left[ c(X_t, \gamma(Y_{[0,t]})) \right] \right| \leq \epsilon/2K$$

for all time stages $t \leq K$ and for all $n > N$. Thus, we have that $\sup_{\gamma \in \Gamma} \left| J_{\beta}(P, T_n, \gamma) - J_{\beta}(P, T, \gamma) \right| \to 0$ as $n \to \infty$.

Now we give the main result of this section.

**Theorem 2.4.3.** Suppose the conditions of Theorem 2.4.2 hold. Then

$$\lim_{n \to \infty} |J_{\beta}^*(P, T_n) - J_{\beta}^*(P, T)| = 0.$$ 

**Proof.** We start with the following bound,

$$|J_{\beta}^*(P, T_n) - J_{\beta}^*(P, T)| \leq \max \left( J_{\beta}(P, T_n, \gamma^*) - J_{\beta}(P, T, \gamma^*), J_{\beta}(P, T, \gamma^*_n) - J_{\beta}(P, T_n, \gamma^*_n) \right)$$

(2.4)

where $\gamma^*$ and $\gamma^*_n$ are the optimal policies respectively for $T$ and $T_n$. Both terms go to 0 by Theorem 2.4.2.

**2.4.3 Absence of Continuity under Setwise Convergence**

We now show that continuity of optimal costs may fail under the setwise convergence of transition kernels. Theorem 2.6.5 in the next section establishes this result for fully observed models. As we note later, a fully observed system can be viewed as a partially observed system with the measurement being the state itself through (2.8), therefore, in
2.4. CONTINUITY OF OPTIMAL COST WITH RESPECT TO CONVERGENCE OF TRANSITION KERNELS (PARTIALLY OBSERVED CASE)

view of space constraints, a separate proof will not be provided for the following result.

**Theorem 2.4.4.** Let $T_n \to T$ setwise. Then, it is not true in general that $J^*_\beta(P, T_n) \to J^*_\beta(P, T)$, even when $X, Y$ and $U$ are compact and $c(x, u)$ is continuous and bounded in $X \times U$.

2.4.4 Continuity under Total Variation

We have the following results.

**Theorem 2.4.5.** Under Assumption 2.3.4 that is

(a) The sequence of transition kernels $T_n$ satisfies the following: $\|T_n(\cdot|x, u_n) - T(\cdot|x, u)\|_{TV} \to 0$ for any sequence $\{u_n\} \subset U$ and $x, u \in X \times U$ such that $u_n \to u$,

(b) The stochastic kernel $T(\cdot|x, u)$ is continuous in total variation in $u$,

(c) The stage-wise cost function $c(x, u)$ is non-negative, bounded and continuous on $U$.

(d) $U$ is compact.

\[ J^*_\beta(P, T_n) \to J^*_\beta(P, T). \]

**Proof.** We start with the following bound,

\[ |J^*_\beta(T_n) - J^*_\beta(T)| \leq \max \left( J^*_\beta(T_n, \gamma^*) - J^*_\beta(T, \gamma^*), J^*_\beta(T, \gamma^*_n) - J^*_\beta(T_n, \gamma^*_n) \right) \]

where $\gamma^*$ and $\gamma^*_n$ are the optimal policies respectively for $T$ and $T_n$.

We now study the following:

\[ \sup_{\gamma \in \Gamma} |J_\beta(P, T_n, \gamma) - J_\beta(P, T, \gamma)| \]
2.4. CONTINUITY OF OPTIMAL COST WITH RESPECT TO CONVERGENCE OF TRANSITION KERNELS (PARTIALLY OBSERVED CASE)

\[ = \sup_{\gamma \in \Gamma} \left| \sum_{t=0}^{\infty} \beta^t \left( E_P^T \left[ c(X_t, \gamma(Y_{[0,t]})) \right] - E_P^{T_n} \left[ c(X_t, \gamma(Y_{[0,t]})) \right] \right) \right| \]

\[ \leq \sum_{t=0}^{\infty} \beta^t \sup_{\gamma \in \Gamma} \left| E_P^T \left[ c(X_t, \gamma(Y_{[0,t]})) \right] - E_P^{T_n} \left[ c(X_t, \gamma(Y_{[0,t]})) \right] \right| \]

In Appendix 2.9.2 we show that for all \( t < \infty \)

\[ \sup_{\gamma \in \Gamma} \left| E_P^T \left[ c(X_t, \gamma(Y_{[0,t]})) \right] - E_P^{T_n} \left[ c(X_t, \gamma(Y_{[0,t]})) \right] \right| \rightarrow 0. \quad (2.5) \]

For any \( \epsilon > 0 \), we choose a \( K < \infty \) such that \( \sum_{i=K+1}^{\infty} \beta^i \| c \|_{\infty} \leq \epsilon/2 \). For the chosen \( K \), we choose an \( N < \infty \) such that

\[ \sup_{\gamma \in \Gamma} \left| E_P^T \left[ c(X_t, \gamma(Y_{[0,t]})) \right] - E_P^{T_n} \left[ c(X_t, \gamma(Y_{[0,t]})) \right] \right| \leq \epsilon/2K \]

for all time stages \( t \leq K \) and for all \( n > N \). Therefore, we have that for any given \( \epsilon > 0 \), for \( n > N \)

\[ \sup_{\gamma \in \Gamma} \left| J_\beta(T_n, \gamma) - J_\beta(T, \gamma) \right| < \epsilon. \quad (2.6) \]

Thus, the result follows. \( \square \)

We now present a result on the rate of convergence. For stochastic control problems, strategic measures are defined [96] as the set of probability measures induced on the product spaces of the state and action pairs by admissible control policies: Given an initial distribution on the state, and a policy, one can uniquely define a probability measure on the infinite product space consistent with finite dimensional distributions, by the Ionescu Tulce theorem [52, Proposition C.10]. Now, define a strategic measure under a policy
2.4. CONTINUITY OF OPTIMAL COST WITH RESPECT TO CONVERGENCE OF TRANSITION KERNELS (PARTIALLY OBSERVED CASE) 37

\( \gamma^n = \{ \gamma^n_0, \gamma^n_1, \ldots, \gamma^n_k, \ldots \} \) as a probability measure defined on \( \mathcal{B}(X \times Y \times U)^{\mathbb{Z}_+} \) by:

\[
P^n_T(d(x_0, y_0, u_0), d(x_1, y_1, u_1), \ldots) = P(dx_0)Q(dy_0|x_0)1_{\{\gamma^n(y_0) \in du_0\}}T(dx_1|x_0, u_0)Q(dy_1|x_1)1_{\{\gamma^n(y_0, y_1) \in du_1\}} \ldots
\]

Next, with uniformity in the total variation convergence, Theorem 2.4.5 is enhanced.

**Theorem 2.4.6.** If the cost function \( c \) is bounded,

\[
|J^*_\beta(P, T_n) - J^*_\beta(P, T)| \leq \|c\|_\infty \frac{\beta}{(\beta - 1)^2} \sup_{x \in X, u \in U} \|T_n(.|x, u) - T(.|x, u)\|_{TV}.
\]

**Proof.** We start with the following bound as before,

\[
|J^*_\beta(T_n, \gamma^*) - J^*_\beta(T, \gamma^*)| \leq \max \left( J^*_\beta(T_n, \gamma^*), J^*_\beta(T_n, \gamma^*_n), J^*_\beta(T, \gamma^*), J^*_\beta(T, \gamma^*_n) \right)
\]

where \( \gamma^* \) and \( \gamma^*_n \) are the optimal policies respectively for \( T \) and \( T_n \).

Then, with \( P^n_T \) and \( P^\gamma_T \) denoting the strategic measures for two chains with a policy \( \gamma \) and kernels \( T_n \) and \( T \), we have

\[
|J^*_\beta(T_n, \gamma) - J^*_\beta(T, \gamma)|
\]

\[
\leq \sum_k \beta^k \|c\|_\infty \|P^n_T(d(x, y, u)_{[0,k]}) - P^\gamma_T(d(x, y, u)_{[0,k]})\|_{TV}
\]
2.5. ROBUSTNESS TO INCORRECT TRANSITION KERNELS (PARTIALLY OBSERVED CASE)

In Appendix 2.9.3 we establish the following relation:

\[ \| P^n_T(d(x, y, u)_{[0,k]}) - P^n_T(d(x, y, u)_{[0,k]}) \|_{TV} \leq k \sup_{x \in X, u \in U} \| T(.|x, u) - T(.|x, u) \|_{TV}. \] 

(2.7)

Using this bound, we will have

\[ \| J^{*}_{\beta}(P, \gamma^*_n) - J^{*}_{\beta}(P, \gamma^*_n) \| \leq \sum_{k} \beta^k \| c \|_\infty k \sup_{x \in X, u \in U} \| T(.|x, u) - T_n(.|x, u) \|_{TV} \]

\[ = \| c \|_\infty \frac{\beta}{(\beta - 1)^2} \sup_{x \in X, u \in U} \| T(.|x, u) - T_n(.|x, u) \|_{TV}. \]

\( \square \)

2.5 Robustness to Incorrect Transition Kernels (Partially Observed Case)

Here, we consider the robustness problem \( \textbf{P2} \): Suppose we design an optimal policy, \( \gamma^*_n \), for a transition kernel, \( T_n \), assuming it is the correct model and apply the policy to the true model whose transition kernel is \( T \). We study the robustness of the sub-optimal policy \( \gamma^*_n \).

2.5.1 Total Variation

Theorem 2.5.1. Suppose the stage-wise cost function \( c(x, u) \) is bounded in \( X \times U \), then

\[ |J_{\beta}(P, T, \gamma^*_n) - J_{\beta}^*(P, T)| \leq 2 \| c \|_\infty \frac{\beta}{(\beta - 1)^2} \sup_{x \in X, u \in U} \| T(.|x, u) - T_n(.|x, u) \|_{TV} \]

for a fixed prior distribution \( P \in \mathcal{P}(X) \), where \( \gamma^*_n \) is the optimal policy designed for the transition kernel \( T_n \).
2.5. ROBUSTNESS TO INCORRECT TRANSITION KERNELS (PARTIALLY OBSERVED CASE)

Proof. We begin with the following,

\[ |J_\beta(\mathcal{T}, \gamma_n^*) - J_\beta^*(\mathcal{T})| \leq |J_\beta(\mathcal{T}, \gamma_n^*) - J_\beta(\mathcal{T}_n, \gamma_n^*)| + |J_\beta(\mathcal{T}_n, \gamma_n^*) - J_\beta(\mathcal{T}, \gamma^*)| \]

The second term is bounded using Theorem 2.4.6. For the first term, we use the proof of Theorem 2.4.6 where we showed that for any \( \gamma \in \Gamma \)

\[ |J_\beta(\mathcal{T}, \gamma) - J_\beta(\mathcal{T}_n, \gamma)| \leq \|c\|_\infty \frac{\beta}{(\beta - 1)^2} \sup_{x \in \mathcal{X}, u \in \mathcal{U}} \|\mathcal{T}(., |x, u) - \mathcal{T}_n(., |x, u)\|_{TV}. \]

Thus, the result follows. \( \square \)

The next theorem gives an asymptotic robustness result.

Theorem 2.5.2. Under Assumption 2.3.4 that is

(a) The sequence of transition kernels \( \mathcal{T}_n \) satisfies the following: \( \|\mathcal{T}_n(., |x, u_n) - \mathcal{T}(., |x, u)\|_{TV} \to 0 \) for any sequence \( \{u_n\} \subset \mathcal{U} \) and \( x, u \in \mathcal{X} \times \mathcal{U} \) such that \( u_n \to u \),

(b) The stochastic kernel \( \mathcal{T}(., |x, u) \) is continuous in total variation in \( u \),

(c) The stage-wise cost function \( c(x, u) \) is non-negative, bounded and continuous on \( \mathcal{U} \).

(d) \( \mathcal{U} \) is compact.

\[ |J_\beta(P, \mathcal{T}, \gamma_n^*) - J_\beta^*(P, \mathcal{T})| \to 0 \]

where \( \gamma_n^* \) is the optimal policy designed for the kernel \( \mathcal{T}_n \).

Proof. We write the following;

\[ |J_\beta(P, \mathcal{T}, \gamma_n^*) - J_\beta^*(P, \mathcal{T})| \leq |J_\beta(P, \mathcal{T}, \gamma_n^*) - J_\beta^*(P, \mathcal{T}_n)| + |J_\beta^*(P, \mathcal{T}_n) - J_\beta^*(P, \mathcal{T})|. \]
2.5. ROBUSTNESS TO INCORRECT TRANSITION KERNELS (PARTIALLY OBSERVED CASE)

The second term goes to 0 by Theorem 2.4.5 and the first term goes to 0 using (2.6) again from the proof of Theorem 2.4.5.

2.5.2 Setwise Convergence

Theorem 2.6.7 in the next section establishes the lack of robustness under setwise convergence of kernels. As we note later, a fully observed system can be viewed as a partially observed system with the measurement being the state itself, see (2.8).

2.5.3 Weak Convergence

**Theorem 2.5.3.** Under Assumptions 2.3.1 and 2.3.2 that is

- The sequence of transition kernels $\mathcal{T}_n$ satisfies the following: $\{\mathcal{T}_n(\cdot|x_n, u_n), n \in \mathbb{N}\}$ converges weakly to $\mathcal{T}(\cdot|x, u)$ for any sequence $\{x_n, u_n\} \subset X \times U$ and $x, u \in X \times U$ such that $(x_n, u_n) \to (x, u)$,

- The stochastic kernel $\mathcal{T}(\cdot|x, u)$ is weakly continuous in $(x, u)$,

- The observation channel $Q(\cdot|x)$ is continuous in total variation i.e., if $x_k \to x$, then $Q(\cdot|x_k) \to Q(\cdot|x)$ in total variation (only for partially observed models),

- The stage-wise cost function $c(x, u)$ is non-negative, bounded and continuous on $X \times U$.

- $U$ is compact.

$|J_\beta(\mathcal{T}, \gamma^*_n) - J^*_\beta(\mathcal{T})| \to 0$, where $\gamma^*_n$ is the optimal policy designed for the transition kernel $\mathcal{T}_n$. 

2.5. ROBUSTNESS TO INCORRECT TRANSITION KERNELS (PARTIALLY OBSERVED CASE)

Proof. We write

$$|J_\beta(\mathcal{T}, \gamma_n^*) - J_\beta^*(\mathcal{T})| \leq |J_\beta(\mathcal{T}, \gamma_n^*) - J_\beta(\mathcal{T}_n, \gamma_n^*)| + |J_\beta(\mathcal{T}_n, \gamma_n^*) - J_\beta(\mathcal{T}, \gamma^*)|.$$ 

The first term goes to 0 by Theorem 2.4.2. For the second term we use Theorem 2.4.3. 

Remark 2.5.1. In this section we study the case where the channel is known to the controller; that is, the true channel model $Q$ is available to the controller. For the case where this is no longer true, the following analysis can be made. If the transition kernel $\mathcal{T}$ and the channel $Q$ are not known, the controller would have an approximating sequence $\mathcal{T}_nQ_n(x_{t+1}, y_{t+1} \in \cdot \times \cdot | x_t, u_t) \in \mathcal{P}(X \times Y)$ for the true joint measure $\mathcal{T}Q(x_{t+1}, y_{t+1} \in \cdot \times \cdot | x_t, u_t) \in \mathcal{P}(X \times Y)$ for all $(x_t, u_t)$. The question then becomes analyzing the convergence of $\mathcal{T}_nQ_n \to \mathcal{T}Q$. Due to space constraints, we do not present explicit results on this problem, however, we note that in [115], a similar joint convergence is studied for convergence of measurement channels and fixed model/prior distributions. The reader can refer to [115, Lemma 2.2] for an analysis on the convergence of $PQ_n \to PQ$, for a single stage problem and for a multi-stage problem, the following can be considered: With

$$|J_\beta^*(\mathcal{T}_nQ_n) - J_\beta^*(\mathcal{T}Q)| \leq |J_\beta^*(\mathcal{T}_nQ_n) - J_\beta(\mathcal{T}Q_n)| + |J_\beta(\mathcal{T}Q_n) - J_\beta^*(\mathcal{T}Q)|,$$

[115, Theorem 6.2] presents sufficient conditions to guarantee the convergence of the second term above. For the first term, the total variation convergence results in this section provide an analysis on the uniform convergence over a class of channels, thus establishing positive results on continuity under the joint convergence of both transition kernels and measurement channels.
2.6 Continuity and Robustness in the Fully Observed Case

In this section, we consider the fully observed case where the controller has direct access to the state variables. We present the results for this case separately, since here we cannot utilize the regularity properties of measurement channels which allows for stronger continuity and robustness results. Similar to the discussions related to (2.2) that is as the operator defined in (2.2) is a contraction and as it admits a fixed point (value function), under measurable selection conditions due to weak or strong (setwise) continuity of transition kernels [52, Section 3.3], for infinite horizon discounted cost problems, optimal policies can be selected from those which are stationary and deterministic. Therefore we will restrict the policies to be stationary and deterministic so that $U_t = \gamma(X_t)$ for some measurable function $\gamma$. Notice also that fully observed models can be viewed as partially observed with the measurement channel thought to be

$$Q(\cdot|x) = \delta_x(\cdot), \quad (2.8)$$

which is only weakly continuous, thus it does not satisfy Assumption [2.3.2].

2.6.1 Weak Convergence

Absence of Continuity under Weak Convergence

We start with a negative result.

**Theorem 2.6.1.** For $T_n \rightarrow T$ weakly, it is not necessarily true that $J^*_\beta(T_n) \rightarrow J^*_\beta(T)$ even when the prior distributions are same and $c(x, u)$ is continuous and bounded in $X \times U$.

**Proof.** We prove the result with a counterexample, similar to the model used in the proof
of Theorem 2.4.1 Let $X = [-1, 1]$, $U = \{-1, 1\}$ and $c(x, u) = (x - u)^2$, the initial distributions are given by $P \sim \delta_1$ that is $X_0 = 1$ and the transition kernels are

$$
T(\cdot | x, u) = \delta_{-1}(x)\left[\frac{1}{2}\delta_1(\cdot) + \frac{1}{2}\delta_{-1}(\cdot)\right] + \delta_1(x)\left[\frac{1}{2}\delta_1(\cdot) + \frac{1}{2}\delta_{-1}(\cdot)\right] + (1 - \delta_{-1}(x))(1 - \delta_1(x))\delta_0(\cdot)
$$

$$
T_n(\cdot | x, u) = \delta_{-1}(x)\left[\frac{1}{2}\delta_{1-1/n}(\cdot) + \frac{1}{2}\delta_{-1+1/n}(\cdot)\right] + \delta_1(x)\left[\frac{1}{2}\delta_{1-1/n}(\cdot) + \frac{1}{2}\delta_{-1+1/n}(\cdot)\right] + (1 - \delta_{-1}(x))(1 - \delta_1(x))\delta_0(\cdot).
$$

It can be seen that $T_n \to T$ weakly according to Definition 2.2.1(i). Under this setup we can calculate the optimal costs as follows;

$$
J_\beta^*(T_n) = \frac{1}{n^2} + \sum_{k=2}^{\infty} \beta^k = \frac{1}{n^2} + \frac{\beta^2}{1 - \beta},
$$

and $J_\beta^*(T) = 0$. Thus, continuity does not hold.

We now present another counter example emphasizing the importance of continuous convergence in the actions. The following counter example shows that without the continuous convergence and regularity assumptions on the kernel $T$, continuity fails even when $T_n(\cdot | x, u) \to T(\cdot | x, u)$ pointwise (for $x, u$) in total variation (also setwise and weakly) and even when the cost function $c(x, u)$ is continuous and bounded. Notice that this example also holds for setwise and weak convergence.

**Example 2.6.1.** Assume that the kernels are given by

$$
T_n(\cdot | x, u) \sim U([u^n, 1 + u^n])
$$

$$
T(\cdot | x, u) \sim \begin{cases} 
U([0, 1]) & \text{if } u \neq 1 \\
U([1, 2]) & \text{if } u = 1
\end{cases}
$$
where $U = [0, 1]$ and $X = \mathbb{R}$. We note first that $\mathcal{T}_n(\cdot|x, u) \to \mathcal{T}(\cdot|x, u)$ in total variation for every fixed $x$ and $u$.

The cost function is in the following form

\[
c(x, u) = \begin{cases} 
2 & \text{if } x \leq \frac{1}{e} \\
2 - \frac{x - \frac{1}{e}}{0.1} & \text{if } \frac{1}{e} < x \leq 0.1 + \frac{1}{e} \\
1 & \text{if } 0.1 + \frac{1}{e} < x \leq 1 + \frac{1}{e} - 0.1 \\
2 - \frac{1 + \frac{1}{e} - x}{0.1} & \text{if } 1 + \frac{1}{e} - 0.1 < x \leq 1 + \frac{1}{e} \\
2 & \text{if } 1 + \frac{1}{e} < x
\end{cases}
\]

Notice that $c(x, u)$ is a continuous function.

With this setup, $\gamma^*(x) = 0$ is an optimal policy for $\mathcal{T}$ since on $[0, 1]$ interval the induced cost is less than the cost induced on $[1, 2]$ interval. The cost under this policy is

\[
J^*_\beta(\mathcal{T}) = \sum_{t=0}^{\infty} \beta^t \left( 2 \times \frac{1}{e} + \frac{0.3}{2} + 0.9 - \frac{1}{e} \right) = \frac{1}{1 - \beta} \left( 1.05 + \frac{1}{e} \right).
\]

For $\mathcal{T}_n$, $\gamma^*_n(x) = e^{-\frac{x}{n}}$ is an optimal policy for every $n$ as $e^{-\frac{1}{n} \times n} = \frac{1}{e}$ and thus the state is distributed between $\frac{1}{e} < x \leq 1 + \frac{1}{e}$ in which interval the cost is the least. Hence, we can write

\[
\lim_{n \to \infty} J_\beta(\mathcal{T}_n, \gamma^*_n) = \sum_{t=0}^{\infty} \beta^t \left( 0.3 + 1 - 0.2 \right) = \frac{1.1}{1 - \beta} \neq \frac{1}{1 - \beta} \left( 1.05 + \frac{1}{e} \right) = J^*_\beta(\mathcal{T}).
\]

\[\diamond\]
A Sufficient Condition for Continuity under Weak Convergence

We will now establish that if the kernels and the model components have some further regularity, continuity does hold.

The assumptions of the following result are same as the assumptions for the partially observed case (Theorem 2.4.3) except for the assumption on the measurement channel $Q$. For clarity, we present the assumptions here separately since for fully observed models we do not need a measurement channel in the analysis.

**Theorem 2.6.2. Under Assumption 2.3.1**

- The sequence of transition kernels $T_n$ satisfies the following: $\{T_n(\cdot|x_n, u_n), n \in \mathbb{N}\}$ converges weakly to $T(\cdot|x, u)$ for any sequence $\{x_n, u_n\} \subset X \times U$ and $x, u \in X \times U$ such that $(x_n, u_n) \to (x, u)$.

- The stochastic kernel $T(\cdot|x, u)$ is weakly continuous in $(x, u)$.

- The stage-wise cost function $c(x, u)$ is non-negative, bounded and continuous on $X \times U$.

- $U$ is compact,

\[ J_\beta(T_n, \gamma_n^*) \to J_\beta(T, \gamma^*), \text{ for any initial state } x_0, \text{ as } n \to \infty. \]

**Proof.** We will use the successive approximations for an inductive argument.

Recall discounted cost optimality operator $T : C_b(Z) \to C_b(Z)$ from (2.2)

\[ (T(v))(x) = \inf_u (c(x, u) + \beta E[v(x_1)|x_0 = x, u_0 = u]) \]
which is a contraction from $C_b(X)$ to itself under the supremum norm and has a fixed point, the value function.

For the kernel $T$, we will denote the approximation functions by

$$v^k(x) = T(v^{k-1})(x)$$

and for the kernel $T_n$ we will use $v^k_n(x)$ to denote the approximation functions, notice that the operator $T$ also depends on $n$ for the model $T_n$ but we will continue using it as $T$ in what follows.

We wish to show that the approximation functions for $T_n$ continuously converge to the ones for $T$. Then, for the first step of the induction we have

$$v^1(x) = c(x, u^*) \quad v^1_n(x_n) = c(x_n, u^*_n)$$

thus we can write,

$$|v^1(x) - v^1_n(x_n)| \leq \sup_{u \in U} |c(x, u) - c(x_n, u)|$$

since $c \in C_b(X \times U)$ and the action space, $U$, is compact, the first step of the induction holds, i.e. $\lim_{n \to \infty} |v^1(x) - v^1_n(x_n)| = 0$.

For the $k^{th}$ step we have,

$$v^k(x) = T(v^{k-1})(x) = \inf_{u} \left[ c(x, u) + \beta \int_X v^{k-1}(x^1)T(dx^1|x, u) \right]$$

$$v^k_n(x_n) = T(v^{k-1}_n)(x_n) = \inf_{u} \left[ c(x_n, u) + \beta \int_X v^{k-1}_n(x^1)T_n(dx^1|x_n, u) \right].$$
2.6. CONTINUITY AND ROBUSTNESS IN THE FULLY OBSERVED CASE

Note that the assumptions of the theorem satisfy the measurable selection criteria and hence we can choose minimizing selectors ([52, Section 3.3]). If we denote the selectors by \( u^* \) and \( u^*_n \), we can write

\[
|v^k(x) - v^k_n(x_n)| \leq \max \left( \left| c(x, u^*) - c(x_n, u^*_n) \right| + \beta \int_X v^{k-1}(x^1) \mathcal{T}(dx^1|x, u^*) - \int_X v^{k-1}_n(x^1) \mathcal{T}_n(dx^1|x_n, u^*_n) \right).
\]

Hence, we can write

\[
|v^k(x) - v^k_n(x_n)| \leq \sup_{u \in U} \left[ |c(x, u) - c(x_n, u)| + \beta \int_X v^{k-1}(x^1) \mathcal{T}(dx^1|x, u) - \int_X v^{k-1}_n(x^1) \mathcal{T}_n(dx^1|x_n, u) \right].
\]  

the first term goes to 0 as \( c(x, u) \) is continuous in \( x \) uniformly over all \( u \in U \). For the second term we write,

\[
\sup_{u \in U} \left| \int_X v^{k-1}(x^1) \mathcal{T}(dx^1|x, u) - \int_X v^{k-1}_n(x^1) \mathcal{T}_n(dx^1|x_n, u) \right|
\]

\[
\leq \sup_{u \in U} \left| \int_X (v^{k-1}(x^1) - v^{k-1}_n(x^1)) \mathcal{T}_n(dx^1|x_n, u) \right|
\]

\[
+ \sup_{u \in U} \left| \int_X v^{k-1}(x^1) \mathcal{T}(dx^1|x, u) - \int_X v^{k-1}_n(x^1) \mathcal{T}_n(dx^1|x_n, u) \right|
\]

for the first term, by the induction argument for any \( x^1_n \rightarrow x^1, |v^{k-1}(x^1) - v^{k-1}_n(x^1_n)| \rightarrow 0 \) (i.e., we have continuous convergence). We also have that \( \mathcal{T}_n(\cdot|x_n, u) \rightarrow \mathcal{T}(\cdot|x, u) \) weakly uniformly over \( u \in U \) as \( U \) is compact. Therefore, using Lemma 2.9.1, the first term goes to 0. For the second term we again use that \( \mathcal{T}_n(\cdot|x_n, u) \) converges weakly to \( \mathcal{T}(\cdot|x, u) \)
uniformly over \( u \in U \). With an almost identical induction argument it can also be shown that \( v^{k-1}(x^1) \) is continuous in \( x^1 \), thus the second term also goes to 0.

So far, we have showed that for any \( k \in \mathbb{N} \),
\[
\lim_{n \to \infty} |v^k_n(x_n) - v^k(x)| = 0 \quad \text{for any } x_n \to x,
\]
in particular it is also true that \( \lim_{n \to \infty} |v^k_n(x) - v^k(x)| = 0 \) for any \( x \).

As we have stated earlier, it can be shown that the approximation operator, \( T \) is a contractive operator under supremum norm with modulus \( \beta \) and it converges to a fixed point which is the value function. Thus, we have
\[
|J_{\beta}(\mathcal{T}, \gamma^*) - v^k(x)| \leq \|c\|_{\infty} \frac{\beta^k}{1 - \beta}, \quad |J_{\beta}^*(\mathcal{T}_n, \gamma_n^*) - v^k_n(x)| \leq \|c\|_{\infty} \frac{\beta^k}{1 - \beta}.
\]

Combining the results,
\[
|J_{\beta}(\mathcal{T}_n, \gamma_n^*) - J_{\beta}(\mathcal{T}, \gamma^*)| \leq |J_{\beta}(\mathcal{T}_n, \gamma_n^*) - v^k_n(x)| + |v^k_n(x) - v^k(x)| + |J_{\beta}(\mathcal{T}, \gamma^*) - v^k(x)|.
\]

Note that the first and the last term can be made arbitrarily small since (2.10) holds for all \( k \in \mathbb{N} \); the second term goes to 0 with an inductive argument for all \( k \in \mathbb{N} \).

**Absence of Robustness under Weak Convergence**

The following result shows that the conditions that satisfy the continuity are not sufficient for robustness in the fully observed models.

**Theorem 2.6.3.** Suppose \( \mathcal{T}_n(\cdot|x_n, u_n) \to \mathcal{T}(\cdot|x, u) \) weakly for every \( x \in X \) and \( u \in U \) and \((x_n, u_n) \to (x, u)\), then it is not true in general that \( J_{\beta}(\mathcal{T}, \gamma_n^*) \to J_{\beta}(\mathcal{T}, \gamma^*) \), even when \( X \) and \( U \) are compact and \( c(x, u) \) is continuous and bounded in \( X \times U \).

**Proof.** We prove the result with a counter example. Take \( X = [0, 2] \) and \( U = \{0, 1, 2\} \).
Suppose the kernels are given in the following form for \( n \geq 1 \):

\[
T_n(x | x, u) = \delta_1 + \frac{1}{n} + 1 \cdot \mathbb{1}_{\{x \geq 1 + 1/n, u = 1\}} + \delta_1 - \frac{1}{n} \cdot \mathbb{1}_{\{x \geq 1 + 1/n, u = 0\}} + \delta_1 \cdot \mathbb{1}_{\{x \geq 1 + 1/n, u = 2\}} + \delta_1 \cdot \mathbb{1}_{\{1 - 1/n < x < 1 + 1/n\}}
\]

\[
T(x | x, u) = \delta_1 (\cdot).
\]

The cost function is given by:

\[
c(x, u) = \begin{cases} 
(x - 1) \mathbb{1}_{x \geq 1} + 0 \mathbb{1}_{x < 1} & \text{if } u = 0, 1 \\
3 & \text{if } u = 2.
\end{cases}
\]

With this setup, an optimal policy for \( T_n \) when the initial state is \( x = 0 \) is given by;

\[
\gamma^*_n(x) = \begin{cases} 
1 & \text{if } x \leq 1 - 1/n \\
0 & \text{if } x \geq 1 + 1/n \\
2 & \text{otherwise}.
\end{cases}
\]

When the initial state is 0, the cost under this policy is \( J_\beta(T_n, \gamma^*_n) = 0 \), therefore the policy \( \gamma^*_n \) is indeed optimal for \( T_n \). An optimal policy for \( T \) is given by \( \gamma^*(x) = 1 \). Thus, the discounted cost values can be calculated as:

\[
J_\beta(T, \gamma^*)_n = \sum_{t=0}^{\infty} \beta^t E[c(X_t, \gamma^*_n(X_t))] = \sum_{t=0}^{\infty} \beta^t c(1, \gamma^*_n(1)) = \sum_{t=0}^{\infty} \beta^t 3 = \frac{3}{1 - \beta}
\]

\[
J_\beta(T, \gamma^*) = 0.
\]
A Sufficient Condition for Robustness under Weak Convergence

We now present a result that establishes robustness if the optimal policies for every initial point are identical. That is, for every \( n \), \( \gamma^*_n \) is optimal for every \( x_0 \in X \) (under the model \( T_n \)). Notice that in the counter example used for Theorem 2.6.3, \( \gamma^*_n \) is not optimal if the initial point is between \( 1 - 1/n \) and \( 1 + 1/n \). A sufficient condition for this property is that \( \gamma^*_n \) solves the discounted cost optimality equation (DCOE) for every initial point.

A policy \( \gamma^* \in \Gamma \) is solves the discounted cost optimality equation and is optimal if it satisfies

\[
J^*_\beta(T, x) = c(x, \gamma^*(x)) + \beta \int J^*_\beta(T, x_1)T(dx_1|x, \gamma^*(x)).
\]

Thus, a policy is optimal for every initial points if it satisfies the DCOE for all initial points \( x \in X \).

**Theorem 2.6.4.** Under Assumption [2.3.1] that is

- The sequence of transition kernels \( T_n \) satisfies the following: \( \{T_n(\cdot|x_n, u_n), n \in \mathbb{N}\} \) converges weakly to \( T(\cdot|x, u) \) for any sequence \( \{x_n, u_n\} \subset X \times U \) and \( x, u \in X \times U \) such that \( (x_n, u_n) \to (x, u) \),

- The stochastic kernel \( T(\cdot|x, u) \) is weakly continuous in \( (x, u) \),

- The stage-wise cost function \( c(x, u) \) is non-negative, bounded and continuous on \( X \times U \),

- \( U \) is compact.
$J_\beta(\mathcal{T}, \gamma_n^*) \to J_\beta(\mathcal{T}, \gamma^*)$ for any initial point $x_0$ if $\gamma_n^*$ is optimal for any initial point for the kernel $\mathcal{T}_n$.

**Proof.** Define the following operator for $\gamma_n^*$, an optimal policy for $\mathcal{T}_n$,

$$(T_n(v))(x_0) = c(x_0, \gamma_n^*(x_0)) + \beta \int v(x_1) \mathcal{T}(dx_1|x_0, \gamma_n^*(x_0))$$

which is a contraction from $C_b(\mathcal{X})$ to itself under the supremum norm with modulus $\beta$ and has a unique fixed point. One can show that the fixed point is $J_\beta(\mathcal{T}, \gamma_n^*)$.

Now we focus on the discounted optimality equation for $\mathcal{T}_n$ for some initial point $x_0^n$ where $x_0^n \to x_0$:

$$J_\beta^*(\mathcal{T}_n, x_0^n) = c(x_0, \gamma_n^*(x_0^n)) + \beta \int J_\beta^*(\mathcal{T}_n, x_1) \mathcal{T}_n(dx_1|x_0^n, \gamma_n^*(x_0^n)).$$

(2.12)

Since the action space $\mathcal{U}$ is compact, we can find a subsequence $n_k$ such that $\gamma_n^*(x_0^{n_k}) \to u^*$ for some $u^* \in \mathcal{U}$. Taking the limit $k \to \infty$ in (2.12) and using Theorem 2.6.2 (continuity) we get

$$J_\beta^*(\mathcal{T}, x_0) = c(x_0, u^*) + \beta \int J_\beta^*(\mathcal{T}, x_1) \mathcal{T}(dx_1|x_0, u^*).$$

(2.13)

Hence $u^*$ satisfies DCOE for $\mathcal{T}$ and is an optimal action for $x_0$. In particular, any convergent subsequence of $\gamma_n^*(x_0^n)$ converges to an optimal action for $x_0$. With this observation, we claim that $T_n^k(J_\beta^*(\mathcal{T}))(x_n) \to J_\beta^*(\mathcal{T}, x)$ for any fixed $k < \infty$ as $n \to \infty$ for some $x_n \to x$ where $T_n^k$ denotes the operator $T_n$ is applied $k$ consecutive times. To show
this, we follow an inductive approach. For \( k = 1 \), we write

\[
T_n(J^*_\beta(T))(x_n) = c(x_n, \gamma^*_n(x_n)) + \beta \int J^*_\beta(T, x_1)T(dx_1|x_n, \gamma^*_n(x_n)).
\]

Suppose \( \lim T_n(J^*_\beta(T))(x_n) \neq J^*_\beta(T, x) \) so that there exists a subsequence \( n_m \) and an \( \epsilon > 0 \) for which \( |T_{n_m}(J^*_\beta(T))(x_{n_m}) - J^*_\beta(T, x)| > \epsilon \) for all \( m \). Since \( \mathcal{U} \) is compact, there exists a further subsequence \( n_{m'} \) such that \( \gamma_{n_{m'}}^*(x_{n_{m'}}) \to u \) for some \( u \in \mathcal{U} \) and as we observed before \( u \) is an optimal action for \( x \) under the kernel \( T \). Hence

\[
\lim_{m' \to \infty} T_{n_{m'}}(J^*_\beta(T))(x_{n_{m'}}) = \lim_{m' \to \infty} c(x_{n_{m'}}, \gamma^*_{n_{m'}}(x_{n_{m'}})) + \beta \int J^*_\beta(T, x_1)T(dx_1|x_{n_{m'}}, \gamma^*_{n_{m'}}(x_{n_{m'}}))
\]

\[
= c(x, u) + \beta \int J^*_\beta(T, x_1)T(dx_1|x, u) = J^*_\beta(T, x)
\]

where the last step follows from the observation that \( u \) is optimal for \( x \) under \( T \). Thus, we reach a contradiction and can conclude that \( T_n(J^*_\beta(T))(x_n) \to J^*_\beta(T, x) \). Now assume that it also holds for \( k - 1 \) so that \( T^{k-1}_n(J^*_\beta(T))(x_n) \to J^*_\beta(T, x) \). We write

\[
T^k_n(J^*_\beta(T))(x_n) = c(x_n, \gamma^*_n(x_n)) + \beta \int T^{k-1}_n(J^*_\beta(T))(x_1)T(dx_1|x_n, \gamma^*_n(x_n)).
\]

Following a similar contradiction argument with the fact that \( T^{k-1}_n(J^*_\beta(T))(x_n) \to J^*_\beta(T, x) \) and using [70, Theorem 3.5] or [97, Theorem 3.5] (weak convergence with varying functions), we can conclude that

\[
\lim_{n \to \infty} T^k_n(J^*_\beta(T))(x_n) \to J^*_\beta(T, x), \quad \forall k < \infty.
\]
Our next claim is that \( T^n_k(J^*_\beta(\mathcal{T}))(x) \to J_\beta(\mathcal{T}, \gamma^*_n, x) \) as \( k \to \infty \). This is true as \( T_n \) is a contraction with modulus \( \beta \) and \( J_\beta(\mathcal{T}, \gamma^*_n) \) is its unique fixed point. Thus \( T^n_k(J^*_\beta(\mathcal{T}))(x) \to J_\beta(\mathcal{T}, \gamma^*_n, x) \) as \( k \to \infty \) and this convergence is uniform over \( n \) as the contraction rate \( \beta \) does not depend on \( n \).

Now, we write:

\[
J_\beta(\mathcal{T}, \gamma^*_n) - J^*_\beta(\mathcal{T}) \leq |J_\beta(\mathcal{T}, \gamma^*_n) - T^n_k(J^*_\beta(\mathcal{T}))| + |T^n_k(J^*_\beta(\mathcal{T})) - J^*_\beta(\mathcal{T})|.
\]

We can make the first term arbitrarily small by choosing \( k \) large enough uniformly over \( n \). For the chosen \( k \), the second term goes to 0 as \( n \to \infty \) by (2.14).

\[
\square
\]

Some remarks are in order.

**Remark 2.6.1 (Policy exchange vs. action exchange).** For the partially observed case, the proof approach we use makes use of policy exchange (e.g. (2.4)) and for this approach total variation continuity of channel \( Q(\cdot|\cdot|x) \) is a key step to deal with the uniform convergence over policies. As we stated before, the channel for fully observed models can be considered in the form of (2.8) which is only weakly continuous and not continuous in total variation. Thus, it may lead to negative results as in Theorem 2.6.3. However, for the fully observed models we can reach continuity and robustness (Theorem 2.6.2, Theorem 2.6.4) using the value iteration approach. With this approach, instead of exchanging policies and analyzing uniform convergence over all policies, we can exchange control actions (e.g. (2.9)) and analyze uniform convergence over the action space \( U \) by using the discounted optimality operator (2.2). Hence, we are able to show convergence only over optimal policies for the fully observed case, i.e. \( J_\beta(\mathcal{T}_n, \gamma^*_n) \to J_\beta(\mathcal{T}, \gamma^*) \) or \( J_\beta(\mathcal{T}, \gamma^*_n) \to J_\beta(\mathcal{T}, \gamma^*) \) where \( \gamma^*_n \) and
2.6. CONTINUITY AND ROBUSTNESS IN THE FULLY OBSERVED CASE

$\gamma^*$ are optimal policies. Whereas, for partially observed models, regularity of the channel allows us to show convergence over any sequence of policies, i.e. $\sup_{\gamma \in \Gamma} |J_\beta(T_n, Q, \gamma) - J_\beta(T, Q, \gamma)| \to 0$.

**Remark 2.6.2.** As we have discussed in subsection 2.4.2, a partially observed model can be reduced to a fully observed process where the state process (beliefs) becomes probability measure valued. Consider the partially observed models with transition kernels $T_n$ and $T$ (with a channel $Q$) and their corresponding fully observed transition kernels $\eta_n$ and $\eta$: following the discussions and techniques in [35] and Chapter 4, one can show that $\eta_n$ and $\eta$ satisfy the conditions of Theorem 2.6.4 and Theorem 2.6.2 that is $\eta_n(\cdot | z_n, u_n) \to \eta(\cdot | z, u)$ for any $(z_n, u_n) \to (z, u)$ under the following set of assumptions

- $T_n(\cdot | x_n, u_n) \to T(\cdot | x, u)$ for any $(x_n, u_n) \to (x, u)$,
- $Q(\cdot | x)$ is continuous on total variation in $x$.

We remark that these conditions also agree with the conditions presented for continuity and robustness of the partially observed models (Theorem 2.4.3 and Theorem 2.5.3).

**Remark 2.6.3.** It can be shown that if we restrict the set of policies to an equicontinuous family of functions, robustness can also be achieved: Under the conditions of Theorem 2.6.2 in this case, $|J_\beta(T, \gamma^*_n) - J_\beta^*(T)| \to 0$. A short proof for this result can be found in Appendix 2.9.4.

### 2.6.2 Setwise Convergence

**Absence of Continuity under Setwise Convergence**

We give a negative result similar to Theorem 2.4.4.
2.6. CONTINUITY AND ROBUSTNESS IN THE FULLY OBSERVED CASE

Theorem 2.6.5. Let \( \mathcal{T}_n \to \mathcal{T} \) setwise then it is not necessarily true that \( J_{\beta}(\mathcal{T}_n) \to J_{\beta}(\mathcal{T}) \) even when \( c(x, u) \) is continuous and bounded in \( X \times U \).

Proof. See Example 2.6.1.

A Sufficient Condition for Continuity under Setwise Convergence

Theorem 2.6.6. Under Assumption 2.3.3 that is

(a) The sequence of transition kernels \( \mathcal{T}_n \) satisfies the following: \( \{\mathcal{T}_n(\cdot|x, u_n), n \in \mathbb{N}\} \) converges setwise to \( \mathcal{T}(\cdot|x, u) \) for any sequence \( \{u_n\} \subset U \) and \( x, u \in X \times U \) such that \( u_n \to u \),

(b) The stochastic kernel \( \mathcal{T}(\cdot|x, u) \) is setwise continuous in \( u \),

(c) The stage-wise cost function \( c(x, u) \) is non-negative, bounded and continuous on \( U \).

(d) \( U \) is compact.

\( J_{\beta}(\mathcal{T}_n, \gamma^*_n) \to J_{\beta}(\mathcal{T}, \gamma^*) \), for any initial state \( x_0 \), as \( n \to \infty \).

Proof. We use the same value iteration technique that we used to prove Theorem 2.6.2.

We wish to show that the approximation functions for \( \mathcal{T}_n \) converge pointwise to the ones for \( \mathcal{T} \). Then, for the first step of the induction we have

\[
v^1(x) = \inf_u c(x, u), \quad v^1_n(x) = \inf_u c(x, u)\]

thus we can write,

\[
|v^1(x) - v^1_n(x)| = 0.
\]
For step $k$ we have,

$$v^k(x) = \inf_u \left[ c(x, u) + \int_X v^{k-1}(x^1) T(dx^1 | x, u) \right]$$

$$v^k_n(x) = \inf_u \left[ c(x, u) + \int_X v^{k-1}_n(x^1) T_n(dx^1 | x, u) \right] .$$

Note that the assumptions of the theorem satisfy the measurable selection criteria and hence we can choose minimizing selectors ([52, Section 3.3]). If we denote the selectors by $u^*$ and $u^*_n$ we can write

$$|v^k(x) - v^k_n(x)| \leq \max \left( \left[ |c(x, u^*) - c(x, u^*_n)| + \int_X v^{k-1}(x^1) T(dx^1 | x, u^*) - \int_X v^{k-1}_n(x^1) T_n(dx^1 | x, u^*_n) \right], \left[ |c(x, u^*_n) - c(x, u^*_n)| + \int_X v^{k-1}(x^1) T(dx^1 | x, u^*_n) - \int_X v^{k-1}_n(x^1) T_n(dx^1 | x, u^*_n) \right] \right) .$$

For the first term we use [89, Theorem 20]. Since $T_n(\cdot | x, u^*_n) \to T(\cdot | x, u^*)$ setwise and $v^{k-1}_n \to v^{k-1}$ pointwise.

For the second term, we use a contradiction argument. Assume that there exists an $\epsilon > 0$ and some subsequence (identified with $n_k$) such that

$$| \int_X v^{k-1}(x^1) T(dx^1 | x, u^*_n) - \int_X v^{k-1}(x^1) T_n(dx^1 | x, u^*_n) | > \epsilon \quad (2.15)$$

Now, take a further subsequence $u^*_{n_k}$ of this sequence which converges to some $u$ whose existence follows from the compactness of $U$. Notice that along this subsequence $T(\cdot | x, u^*_{n_k}) \to T(\cdot | x, u)$ and $T_{n_k}(\cdot | x, u^*_{n_k}) \to T(\cdot | x, u)$. Thus, using the induction step and [89, Theorem 20] the above term converges to 0 along the subsequence indexed by $n_k$ which contradicts with (2.15). The rest of the proof follows from the arguments in Theorem 5.2. \qed
Absence of Robustness under Setwise Convergence

Now, we give a result showing that even if continuity holds under the setwise convergence of the kernels, robustness may not be satisfied. The counterexample used in the proof builds on [115, Section 3.2.1].

**Theorem 2.6.7.** Suppose $T_n(\cdot|x_n,u_n) \to T(\cdot|x,u)$ setwise for every $x \in X$ and $u \in U$ and $(x_n,u_n) \to (x,u)$, then it is not true in general that $J_\beta(T,\gamma^*_n) \to J_\beta(T,\gamma^*)$, even when $X$ and $U$ are compact and $c(x,u)$ is continuous and bounded in $X \times U$.

**Proof.** We prove the result with a counterexample. Define

$$L_{n,k} = \left[\frac{2k - 2}{2n}, \frac{2k - 1}{2n}\right], \quad R_{n,k} = \left[\frac{2k - 1}{2n}, \frac{k}{n}\right].$$

Let $L = \{y \in \bigcup_{k=1}^n L_{n,k}\}$ and $R = \{y \in \bigcup_{k=1}^n R_{n,k}\}$. Next, define the square-wave function by

$$h_n(t) = 1_{\{t \in L\}} - 1_{\{t \in R\}}.$$

Define two sequences of probability density functions as

$$f_n(t) = (1 + h_n(t))1_{\{t \in [0,1]\}}, \quad g_n(t) = (1 - h_n(t))1_{\{t \in [0,1]\}}$$

Consider the kernels given in the following form for $n \geq 1$:

$$T_n(\cdot|x,u) \sim 1_{\{x \in L, u = 1\}} f_n(\cdot) + 1_{\{x \in L, u = 0\}} g_n(\cdot) + 1_{\{x \in R, u = 0\}} f_n(\cdot) + 1_{\{x \in R, u = 1\}} g_n(\cdot)$$
2.6. CONTINUITY AND ROBUSTNESS IN THE FULLY OBSERVED CASE

\[ T(\cdot|x, u) \sim U([0, 1]). \]

By the proof of Riemann-Lebesgue lemma \[110\] Theorem 12.21

\[ \lim_{n \to \infty} \int_0^1 h_n(t)g(t)\,dt = 0 \text{ for all } g \in L_1([0, 1], \mathbb{R}), \]

and therefore

\[ \lim_{n \to \infty} \int_0^1 f_n(t)g(t)\,dt = \int_0^1 g(t) \text{ for all } g \in L_1([0, 1], \mathbb{R}). \]

As a result, \( T_n(\cdot|x, u) \to T(\cdot|x, u) \) setwise for every \( x \in X \) and \( u \in U \).

The cost function is given by:

\[ c(x, u) = \begin{cases} 2 & \text{if } u = 0 \\ x & \text{if } u = 1. \end{cases} \]

Notice that if the system starts anywhere on \( L \), it does not matter how we define \( \gamma^*_n \) for \( x \in R \) as the state always stays at \( L \). Thus, with this setup, it can be seen that an optimal policy for \( T_n \) is given by:

\[ \gamma^*_n(x) = \begin{cases} 1 & \text{if } x \in L \\ 0 & \text{if } x \in R. \end{cases} \]

when the initial state is \( x = 0 \), which belongs to \( L \) for any \( n \geq 1 \). The optimal policy for
2.6. CONTINUITY AND ROBUSTNESS IN THE FULLY OBSERVED CASE

$T$ is given by $\gamma^*(x) = 1$. The discounted cost values can be calculated as follows:

$$J_{\beta}(T, \gamma^*) = \sum_{t=0}^{\infty} \beta^t E_T[c(X, 1)] = \sum_{t=0}^{\infty} \beta^t \int_0^1 x dx = \frac{1}{2(1-\beta)}.$$  

Building on the calculations in [62 Theorem 2.7], the cost under the policy $\gamma^*_n$ is calculated as:

$$J_{\beta}(T, \gamma^*_n) = \sum_{t=0}^{\infty} \beta^t E_T[c(X, \gamma^*_n(X))] = \frac{1}{1-\beta} \left( \int_L c(x, 1) dx + \int_R c(x, 0) dx \right)$$

$$= \frac{1}{1-\beta} \left( \int_L x dx + \int_R 2 dx \right) = \frac{1}{1-\beta} \left( \frac{1}{4} - \frac{1}{8n} + 1 \right) \to \frac{5}{4(1-\beta)}$$

which completes the proof.

A Sufficient Condition for Robustness under Setwise Convergence

We now present a similar result to Theorem 2.6.4 that is we show that under the conditions of Theorem 2.6.6, if further for every $n$, $\gamma^*_n$ is optimal for every $x_0 \in X$ (under the model $T_n$) then robustness holds under setwise convergence.

**Theorem 2.6.8.** Suppose Assumption 2.3.3 holds, that is

(a) The sequence of transition kernels $T_n$ satisfies the following: $\{T_n(\cdot|x, u_n), n \in \mathbb{N}\}$ converges setwise to $T(\cdot|x, u)$ for any sequence $\{u_n\} \subset U$ and $x, u \in X \times U$ such that $u_n \to u$,

(b) The stochastic kernel $T(\cdot|x, u)$ is setwise continuous in $u$,

(c) The stage-wise cost function $c(x, u)$ is non-negative, bounded and continuous on $U$.

(d) $U$ is compact.
if further we have that for every $n$, $\gamma^*_n$ is optimal for every $x_0 \in X$ (under the model $T_n$) then $J_\beta(T, \gamma^*_n) \to J_\beta(T, \gamma^*)$.

**Proof.** We use the same proof technique as we used for Theorem 2.6.4. Using the properties of setwise convergence one can again show that

$$
\lim_{n \to \infty} T_n^k(J_\beta^*(T))(x_n) \to J_\beta^*(T, x), \quad \forall k < \infty.
$$

(2.16)

Recalling the operator $T_n^k$ from (2.11) ($T_n$ applied $k$ times), it is also true that $T_n^k(J_\beta^*(T))(x) \to J_\beta(T, \gamma^*_n, x)$ as $k \to \infty$ uniformly over $n$. Hence

$$
J_\beta(T, \gamma^*_n) - J_\beta^*(T) \leq |J_\beta(T, \gamma^*_n) - T_n^k(J_\beta^*(T))| + |T_n^k(J_\beta^*(T)) - J_\beta^*(T)| \to 0.
$$

2.6.3 **Total Variation**

The continuity result in Theorem 2.4.5 and the robustness result in Theorem 2.5.1 apply to this case since the fully observed model may be viewed as a partially observed model with the measurement channel $Q$ given in (2.8).

**Remark 2.6.4.** We note that if the action and state spaces are finite, then total variation convergence and weak convergence coincide and thus Theorem 2.4.5, Theorem 2.4.6 and Theorem 2.5.1 from the partially observed case directly apply to this case considering the channel as a perfect channel. Thus, the only assumptions needed to establish continuity and robustness are
2.7. **FINITE MODEL APPROXIMATIONS FOR WEAKLY CONTINUOUS MDPs AS AN INSTANCE OF THE ROBUSTNESS PROBLEM**

- \( T_n(\cdot|x,u) \to T(\cdot|x,u) \) in total variation for all \( x \in X \).
- \( T(\cdot|x,u) \) is continuous in total variation in \( u \) for every given \( x \in X \).
- \( \mathbb{U} \) is compact.

Since the spaces are finite, these set of assumptions reduces to

- \( T_n(\cdot|x,u) \to T(\cdot|x,u) \) in total variation for all \( x \in X \) and \( u \in \mathbb{U} \).

**Remark 2.6.5.** We note that all of the results we present in this section also apply to finite horizon problems. If we define a finite horizon problem by

\[
J(P,T,\gamma) = \sum_{t=0}^{T} E_{P,Q}^T[c(X_t, U_t)]
\]

the continuity and robustness properties hold under the same conditions we have presented for the infinite horizon discounted problem.

**2.7 Finite Model Approximations for Weakly Continuous MDPs as an Instance of the Robustness Problem**

We now discuss whether an approximation of an MDP model with a standard Borel space with a finite MDPs can be viewed as an instance of robustness problem to incorrect models and whether our results can be applied. We first review the following due to Saldi et. al. [93].

**2.7.1 Review of Finitely Quantized Approximations to Standard Borel MDPs**

Consider an MDP which is quantized as follows.
Finite State Approximate MDP: Quantization of the State Space Let $d_{\mathcal{X}}$ denote the metric on $\mathcal{X}$. For each $n \geq 1$, there exists a finite subset \(\{x_{n,i}\}_{i=1}^{k_n}\) of $\mathcal{X}$ such that

$$\min_{i \in \{1, \ldots, k_n\}} d_{\mathcal{X}}(x, x_{n,i}) < 1/n$$

for all $x \in \mathcal{X}$.

Let $\mathcal{X}_n := \{x_{n,1}, \ldots, x_{n,k_n}\}$ and define $Q_n$ mapping any $x \in \mathcal{X}$ to the nearest element of $\mathcal{X}_n$, i.e.,

$$Q_n(x) := \arg \min_{x_{n,i} \in \mathcal{X}_n} d_{\mathcal{X}}(x, x_{n,i}).$$

For each $n$, a partition $\{S_{n,i}\}_{i=1}^{k_n}$ of the state space $\mathcal{X}$ is induced by $Q_n$ by setting

$$S_{n,i} = \{x \in \mathcal{X} : Q_n(x) = x_{n,i}\}.$$

Let $\psi$ be a probability measure on $\mathcal{X}$ which satisfies

$$\psi(S_{n,i}) > 0 \text{ for all } i, n,$$

and define probability measures $\psi_{n,i}$ on $S_{n,i}$ by restricting $\psi$ to $S_{n,i}$:

$$\psi_{n,i}(\cdot) := \frac{\psi(\cdot)}{\psi(S_{n,i})}.$$

Using $\{\psi_{n,i}\}$, we define a sequence of finite-state MDPs, denoted as $f$-MDP$_m$, to approximate the compact-state MDP.

For each $m$, $f$-MDP$_m$ is defined as: $(\mathcal{X}_n, \mathcal{U}, T_n, c_n)$, and the one-stage cost function
2.7. FINITE MODEL APPROXIMATIONS FOR WEAKLY CONTINUOUS MDPS
AS AN INSTANCE OF THE ROBUSTNESS PROBLEM

$c_n : \mathcal{X}_n \times \mathbb{U} \to [0, \infty)$ and the transition probability $\mathcal{T}_n$ on $\mathcal{X}_n$ given $\mathcal{X}_n \times \mathbb{U}$ are given by

$$c_n(x_{n,i}, a) := \int_{\mathbb{S}_{n,i}} c(x, a) \psi_{n,i}(dx)$$

$$\mathcal{T}_n(\cdot | x_{n,i}, a) := \int_{\mathbb{S}_{n,i}} Q_n * \mathcal{T}(\cdot | x, a) \psi_{n,i}(dx),$$

where $Q_n * \mathcal{T}(\cdot | x, a) \in \mathcal{P}(\mathcal{X}_n)$ is the pushforward of the measure $\mathcal{T}(\cdot | x, a)$ with respect to $Q_n; that is,

$$Q_n * \mathcal{T}(z_{n,j}|x, a) = \mathcal{T}(\{y \in \mathcal{X} : Q_n(y) = x_{n,j}\}|x, a),$$

for all $x_{n,j} \in \mathcal{X}_n$.

**Finite Action Approximate MDP: Quantization of the Action Space**

Let $d_\mathbb{U}$ denote the metric on $\mathbb{U}$. Since the action space $\mathbb{U}$ is compact and thus totally bounded, one can find a sequence of finite sets $\Lambda_n = \{a_{n,1}, \ldots, a_{n,k_n}\} \subset \mathbb{U}$ such that for all $n$,

$$\min_{i \in \{1, \ldots, k_n\}} d_\mathbb{U}(a, a_{n,i}) < 1/n \text{ for all } a \in \mathbb{U}.$$

In other words, $\Lambda_n$ is a $1/n$-net in $\mathbb{U}$. Let us assume that the sequence $\{\Lambda_n\}_{n \geq 1}$ is fixed. To ease the notation in the sequel, let us define the mapping $\Upsilon_n$

$$\Upsilon_n(f)(x) := \arg \min_{a \in \Lambda_n} d_\mathbb{U}(f(x), a), \quad (2.17)$$

where ties are broken so that $\Upsilon_n(f)(x)$ is measurable.

It is known that finite quantization policies are nearly optimal under the conditions to be presented below, see [92, Theorem 3.2]. Thus, to make the presentation shorter, we will either assume that the action set is finite, or it has been approximated by a finite
action space through the construction above, whose near optimality has been shown in \[92\]. Assuming finite action sets will help us avoid measurability issues (see universal measurability discussions in \[93\]) as well as issues with existence of optimal policies.

**Assumption 2.7.1.**

(a) The one stage cost function $c$ is nonnegative and continuous.

(b) The stochastic kernel $T(\cdot|x,a)$ is weakly continuous in $(x,a) \in X \times U$.

(c) $U$ is finite.

(d) $X$ is compact.

We note that condition (d) in Assumption 2.7.1 as presented in \[93\] was more general, but we have used the simpler version here for clarity in exposition.

One can write the following fixed point equation for the finite MDP

$$J_n^\beta(x) = \min_{a \in U} \left\{ c_n(x,a) + \beta \sum_{x_1 \in X_n} J_n^\beta(x_1)T_n(x_1|x,a) \right\}$$

where $T_n$ is the transition model for the finite MDP and $c_n$ is the cost function defined on the finite model. Since the action space is finite, we can find an optimal policy, say $f^*_n$ for this fixed point equation. One can also simply extend $J_n^\beta$ and $f^*_n$, which are defined on $X_n$ to the entire state space $X$ by taking them constant over the quantization bins $S_{n,i}$. If we call the extended versions $\hat{J}_n^\beta$ and $\hat{f}_n$, the following result can be established:

**Theorem 2.7.1.** \[93\] Theorem 2.2 and 4.1] Suppose Assumption 2.7.1 holds. Then, for any $\beta \in (0,1)$ the discounted cost of the deterministic stationary policy $\hat{f}_n$, obtained by extending the discounted optimal policy $f^*_n$ of $f$-MDP$_m$ to $X$ (i.e., $\hat{f}_n = f^*_n \circ Q_n$), converges
to the discounted value function $J^*$ of the compact-state MDP:

$$
\lim_{n \to \infty} \| \hat{J}^n_\beta (\cdot) - J^*_\beta (\cdot) \| = 0 \quad \text{and} \quad \lim_{n \to \infty} \| J_\beta (\hat{f}_n, \cdot) - J^*_\beta \| = 0. \quad (2.18)
$$

Theorems 2.7.1 shows that under Assumption 2.7.1 an optimal solution can be approximated via the solutions of finite models.

We now show that the above approximation scheme can be viewed in relation to our robustness results.

**Proof of Theorem 2.7.1 via Results from Section 2.6.** With the introduced setup, one can see that the extended value function and optimal policy for the finite model satisfy the following:

$$
\hat{J}^n_\beta (x) = \min_{a \in U} \left\{ \hat{c}_n (x, u) + \beta \int \hat{J}^n_\beta (x_1) \hat{T}_n (dx_1 | x, u) \right\}
$$

where $\hat{c}_n$ is the extended version of $c_n$ to the state space $\mathcal{X}$ by making it constant over the quantization bins $\{S_{n,i}\}_i$ and $\hat{T}_n$ is such that for any function $f$

$$
\int f(x_1) \hat{T}_n (dx_1 | x, u) := \int_{x_1 \in \mathcal{X}} \int_{z \in S_{n,i}} f(x_1) T(dx_1 | z, u) \psi_{n,i} (dz)
$$

where $S_{n,i}$ is the quantization bin that $x$ belongs to.

With this setup, one can see that for any $x_n \to x$ we have $\hat{c}_n (x_n, u) \to c(x, u)$. Using the fact that $\psi_{n,i}$ is a probability measure on $S_{n,i}$ and that $\int f(x_1) T(dx_1 | x, u)$ is a constant under integration over $\psi_{n,i} (dz)$, for any continuous and bounded $f$ we can write

$$
\int f(x_1) \hat{T}_n (dx_1 | x, u) - \int f(x_1) T(dx_1 | x, u)
$$
2.8. EMPIRICAL MODEL LEARNING AS AN INSTANCE OF THE ROBUSTNESS PROBLEM

\[ = \int_{z \in \mathcal{S}_{n,i}} \int_{x_1 \in \mathcal{X}} f(x_1) T(dx_1|z, u) \psi_{n,i}(dz) - \int_{z \in \mathcal{S}_{n,i}} \int_{x_1 \in \mathcal{X}} f(x_1) T(dx_1|x, u) \psi_{n,i}(dz) \]
\[ \leq \int_{z \in \mathcal{S}_{n,i}} \left| \int_{x_1 \in \mathcal{X}} f(x_1) T(dx_1|z, u) - \int_{x_1 \in \mathcal{X}} f(x_1) T(dx_1|x, u) \right| \psi_{n,i}(dz). \]

Since, \( T(\cdot|x, u) \) is weakly continuous in \( x \), for any given \( \epsilon \), there exists a \( \delta > 0 \) such that \( |x - y| \leq \delta \) implies \( | \int f(x_1) T(dx_1|y, u) - \int f(x_1) T(dx_1|x, u) | \leq \epsilon. \)

We know fix an \( \epsilon \) and choose such a \( \delta \). We need to show that we can find an \( N < \infty \) such that for all \( n > N \), if we denote the bin \( x_n \) belongs to by \( \mathcal{S}_{n,i} \), then for any \( z \in \mathcal{S}_{n,i} \), we must have \( |z - x| \leq \delta \). Since \( x_n \to x \), we can find \( N \) with \( |x_n - x| \leq \delta/2 \) for all \( n > N \).

Furthermore, by construction \( |z - x_n| \leq 1/n \), hence by choosing \( N > 2/\delta \), we can make \( |z - x_n| \leq \delta/2 \). Then, by choosing the larger \( N \), we can have \( |z - x| \leq |z - x_n| + |x_n - x| \leq \delta \). Hence, we can conclude that

\[ \int f(x_1) \hat{T}_n(dx_1|x_n, u) \to \int f(x_1) T(dx_1|x, u). \]

Then, Assumption 2.3.1 holds under Assumption 2.7.1 and we can conclude the proof using Theorem 2.6.4 and Theorem 2.6.2

2.8 Empirical Model Learning as an Instance of the Robustness Problem

For unknown probability measures, a common practice is to learn them via test inputs or empirical observations. Let \( \{(X_i), i \in \mathbb{N}\} \) be an \( \mathcal{X} \)-valued i.i.d random variable sequence generated according to some distribution \( \mu \).

Defining for every (fixed) Borel \( B \subset \mathcal{X} \), and \( n \in \mathbb{N} \), the empirical occupation measures

\[ \mu_n(B) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_i \in B\}}, \]
one has $\mu_n(B) \to \mu(B)$ almost surely (a.s.) by the strong law of large numbers. Also, $\mu_n \to \mu$ weakly with probability one ([31], Theorem 11.4.1).

However, $\mu_n$ can not converge to $\mu$ in total variation, in general. On the other hand, if we know that $\mu$ admits a density, we can find estimators to estimate $\mu$ under total variation [28].

As discussed above, the empirical averages converge almost surely. By a similar reasoning, for a given bounded measurable function $f$, $\int \mu_n(dx) f(x)$ converges to $\int \mu(dx) f(x)$. This then also holds for any countable collection of functions, $f_1, f_2, \cdots$. A relevant question is the following: Can one ensure uniform convergence (over a family of functions) with arbitrary precision by only guaranteeing convergence for a finite or countably infinite collection of functions? This entails the problem of covering a family of functions with arbitrarily small neighborhoods of finitely many functions under an appropriate distance metric. The answer to this question is studied by the theory of empirical risk minimization: In the learning theoretic context when one tries to estimate the source distribution, the convergence of optimal costs under $\mu_n$ to the cost optimal for $\mu$ is called the consistency of empirical risk minimization [107]. In particular, if the following uniform convergence holds,

$$
\lim_{n \to \infty} \sup_{f \in \mathcal{F}} \left| \int f(x) \mu_n(dx) - \int f(x) \mu(dx) \right| = 0,
$$

for a class of measurable functions $\mathcal{F}$, then $\mathcal{F}$ is called a $\mu$-Glivenko-Cantelli class [32]. If the class $\mathcal{F}$ is $\mu$-Glivenko-Cantelli for every $\mu$, it is called a universal Glivenko-Cantelli class. One example of a universal Glivenko-Cantelli family of real functions on $\mathbb{R}^N$ is the family $\{f : \|f\|_{BL} \leq M\}$ for some $0 < M < \infty$, where $\|f\|_{BL} = \|f\|_{\infty} + \sup_{x_1 \neq x_2} \left| \frac{f(x_1) - f(x_2)}{x_1 - x_2} \right|$ ([32]). For related characterizations and further examples, see [87, 106, 30].
For our analysis, we will mainly make use of the fact that the empirical occupation measures $\mu_n$ converge weakly almost surely to the true underlying measure $\mu$.

The above discussion is for the i.i.d. observations and for fixed probability measures. In the case of unknown dynamics or unknown transition models, for every different $(x, u) \in X \times U$ pair, $T(\cdot|x,u)$ is a different probability measure, and hence there are possibly uncountably many unknown probability measures. Thus, in this section we focus on some special settings.

Let $T(\cdot|x,u)$ be a transition kernel given previous state and action variables $x \in X, u \in U$, which is unknown to the decision maker (DM). Suppose the DM builds a model for the transition kernels, $T_n(\cdot|x,u)$, for all possible $x \in X, u \in U$ by collecting training data $\{x_t, u_t, t \leq n\}$ from the evolving system.

The question we are interested in is that, do we have that the cost calculated under $T_n$ converges to the true cost (i.e., do we have that the cost obtained from applying the optimal policy for the empirical model converges to the true cost as the training length increases)? We will refer to this property as being asymptotically robust under empirical learning.

**Corollary 2.8.1** (to Theorem 2.4.6 and Theorem 2.5.1). Suppose we are given the following dynamics for finite state space, $X$, and finite action space, $U$,

$$x_{t+1} = f(x_t, u_t, w_t), \quad y_t = g(x_t, v_t)$$

where $\{w_t\}$ and $\{v_t\}$ are i.i.d. noise processes and the noise models are unknown. Suppose that there is an initial training period so that under some policy, every $x, u$ pair is visited infinitely often if training were to continue indefinitely, but that the training ends at some finite time. Let us assume that, through this training, we empirically learn the transition
dynamics such that for every (fixed) Borel $B \subset X$, for every $x \in X$, $u \in U$ and $n \in \mathbb{N}$, the empirical occupation measures are

$$
T_n(B|x_0 = x, u_0 = u) = \frac{\sum_{i=1}^{n} 1\{x_i \in B, x_{i-1} = x, u_{i-1} = u\}}{\sum_{i=1}^{n} 1\{x_{i-1} = x, u_{i-1} = u\}}.
$$

Then we have that $J_n^*(T_n) \to J^*(T)$ and $J_n^*(T, \gamma_n^*) \to J_n^*(T)$, where $\gamma_n^*$ is the optimal policy designed for $T_n$. Since the channel model $g$ has no restrictions, this result also applies to the fully observed model setup by taking $g(x_t, v_t) = x_t$.

**Proof.** We have that $T_n(\cdot|x, u) \to T(\cdot|x, u)$ weakly for every $x \in X$, $u \in U$ almost surely by law of large numbers. Since the spaces are finite, we also have $T_n(\cdot|x, u) \to T(\cdot|x, u)$ under total variation. By Theorem 2.4.6 and Theorem 2.5.1 the results follow. 

The following holds for more general spaces.

**Corollary 2.8.2** (to Theorem 2.5.3, Theorem 2.4.3, Theorem 2.6.2 and Theorem 2.6.4).

Suppose we are given the following dynamics with state space $X$ and action space $U$,

$$
x_{t+1} = f(x_t, u_t, w_t), \quad y_t = g(x_t, v_t),
$$

where $\{w_t\}$ and $\{v_t\}$ are i.i.d. noise processes and the noise models are unknown. Suppose that $f(x, u, \cdot) : W \to X$ is invertible for all fixed $(x, u)$ and $f(x, u, w)$ is continuous and bounded on $X \times U \times W$. We construct the empirical measures for the noise process $w_t$ such that for every (fixed) Borel $B \subset W$, and for every $n \in \mathbb{N}$, the empirical occupation measures are

$$
\mu_n(B) = \frac{1}{n} \sum_{i=1}^{n} 1\{f_{i-1}^{-1}(x_{i-1}, u_{i-1})(x_i) \in B\}
$$

(2.20)
where \( f_{x_{i-1},u_{i-1}}^{-1}(x_i) \) denotes the inverse of \( f(x_{i-1}, u_{i-1}, w) : \mathbb{W} \to \mathbb{X} \) for given \((x_{i-1}, u_{i-1})\).

Using the noise measurements, we construct the empirical transition kernel estimates for any \((x_0, u_0)\) and Borel \(B\) as

\[
\mathcal{T}_n(B|x_0, u_0) = \mu_n(f_{x_0,u_0}^{-1}(B)).
\]

(i) If the measurement channel (represented by the function \( g \)) is continuous in total variation then \( J_\beta^*(\mathcal{T}_n) \to J_\beta^*(\mathcal{T}) \) and \( J_\beta(\mathcal{T}, \gamma_n^*) \to J_\beta^*(\mathcal{T}) \), where \( \gamma_n^* \) is the optimal policy designed for \( \mathcal{T}_n \) for all initial points.

(ii) If the measurement channel is in the form \( g(x_t, v_t) = x_t \) (i.e. fully observed) then \( J_\beta^*(\mathcal{T}_n) \to J_\beta^*(\mathcal{T}) \) and if further for every \( n \), \( \gamma_n^* \) is optimal for every \( x_0 \in \mathbb{X} \) (under the model \( \mathcal{T}_n \)) then \( J_\beta(\mathcal{T}, \gamma_n^*) \to J_\beta^*(\mathcal{T}) \).

Proof. We have \( \mu_n \to \mu \) weakly with probability one where \( \mu \) is the model. We claim that the transition kernels are such that \( \mathcal{T}_n(\cdot|x_n, u_n) \to \mathcal{T}(\cdot|x, u) \) weakly for any \((x_n, u_n) \to (x, u)\). To see that observe the following for \( h \in C_b(\mathbb{X}) \)

\[
\int h(x_1)\mathcal{T}_n(dx_1|x_n, u_n) - \int h(x_1)\mathcal{T}(dx_1|x, u) \\
= \int h(f(x_n, u_n, w))\mu_n(dw) - \int h(f(x, u, w))\mu(dw) \to 0,
\]

where \( \mu_n \) is the empirical measure for \( w_t \) and \( \mu \) is the true measure again. For the last step, we used that \( \mu_n \to \mu \) weakly and \( h(f(x_n, u_n, w)) \) continuously converge to \( h(f(x, u, w)) \) i.e. \( h(f(x_n, u_n, w_n)) \to h(f(x, u, w)) \) for some \( w_n \to w \) since \( f \) and \( g \) are continuous functions. Similarly, it can be also shown that \( \mathcal{T}_n(\cdot|x, u) \) and \( \mathcal{T}(\cdot|x, u) \) are weakly continuous on \((x, u)\). Thus, for the case where the channel is continuous in total variation by Theorem
2.5.3 and Theorem 2.4.3 if $c(x, u)$ is bounded and $U$ is compact the result follows.

For the fully observed case, $J^*_{\beta}(T_n) \rightarrow J^*_{\beta}(T)$ by Theorem 2.6.2 and $J_{\beta}(T, \gamma^*_n) \rightarrow J^*_{\beta}(T)$ by Theorem 2.6.4.

Now, we give a similar result with the assumption that the noise process of the dynamics admits a continuous probability density function.

**Corollary 2.8.3** (to Theorem 2.4.6 and Theorem 2.5.1). Suppose we are given the following dynamics for real vector state space $X$ and action space $U$

$$x_{t+1} = f(x_t, u_t, w_t), \quad y_t = g(x_t, v_t),$$

where $\{w_t\}$ and $\{v_t\}$ are i.i.d. noise processes and the noise models are unknown but it is known that the noise $w_t$ admits a continuous probability density function. Suppose that $f(x, u, \cdot) : W \rightarrow X$ is invertible for all $(x, u)$. We collect i.i.d. samples of $\{w_t\}$ as in (2.20) and use them to construct an estimator, $\tilde{\mu}_n$, as described in [28] which consistently estimates $\mu$ in total variation. Using these empirical estimates, we construct the empirical transition kernel estimates for any $(x_0, u_0)$ and Borel $B$ as

$$T_n(B|x_0, u_0) = \tilde{\mu}_n(f^{-1}_{x_0,u_0}(B)).$$

Then independent of the channel, $J^*_{\beta}(T_n) \rightarrow J^*_{\beta}(T)$ and $J_{\beta}(T, \gamma^*_n) \rightarrow J^*_{\beta}(T)$, where $\gamma^*_n$ is the optimal policy designed for $T_n$. Since the channel model $g$ has no restrictions, this result also applies to the fully observed model setup by taking $g(x_t, v_t) = x_t$.

**Proof.** By [28] we can estimate $\mu$ in total variation so that almost surely $\lim_{n \rightarrow \infty} \|\tilde{\mu}_n - \mu\|_{TV} = 0$. We claim that the convergence of $\tilde{\mu}_n$ to $\mu$ under total variation metric implies
2.8. EMPIRICAL MODEL LEARNING AS AN INSTANCE OF THE ROBUSTNESS PROBLEM

the convergence of $T_n$ to $T$ in total variation uniformly over all $x \in X$ and $u \in U$ i.e.
\[
\lim_{n \to \infty} \sup_{x,u} \| T_n(\cdot|x,u) - T(\cdot|x,u) \|_{TV} = 0.
\]
Observe the following:

\[
\sup_{x,u} \| T_n(\cdot|x,u) - T(\cdot|x,u) \|_{TV} = \sup_{x,u} \sup_{||h||_\infty \leq 1} \left| \int h(x_1) T_n(dx_1|x,u) - \int h(x_1) T(dx_1|x,u) \right|
\]
\[
= \sup_{x,u} \sup_{||h||_\infty \leq 1} \left| \int h(f(x,u,w)) \tilde{\mu}_n(dw) - \int h(f(x,u,w)) \mu(dw) \right| \leq \| \tilde{\mu}_n - \mu \|_{TV} \to 0.
\]

Thus, by Theorem 2.4.6 and Theorem 2.5.1, the result follows.

The following example presents some system and channel models which satisfy the requirements of the above corollaries.

**Example 2.8.1.** Let $X, Y, U$ be finite dimensional real vector spaces with

\[
x_{t+1} = f(x_t, u_t) + w_t, \quad y_t = h(x_t, v_t)
\]

for unknown i.i.d. noise processes $\{w_t\}$ and $\{v_t\}$.

(i) Suppose the channel is in the following form; $y_t = h(x_t, v_t) = x_t + v_t$ where $v_t$ admits a density (e.g. Gaussian density). It can be shown by an application of Scheffé’s theorem that the channels in this form are continuous in total variation. If further $f(x_t, u_t)$ is continuous and bounded then the requirements of Corollary 2.8.2 hold for partially observed models.

(ii) If the channel is in the following form; $x_t = h(x_t, v_t)$ then the system is fully observed. If further $f(x_t, u_t)$ is continuous and bounded then the requirements of Corollary 2.8.2 holds for fully observed models.

(iii) Suppose the function $f(x_t, u_t)$ is known, if the noise process $w_t$ admits a continuous
density, then one can estimate the noise model in total variation in a consistent way (see \cite{28}). Hence, the conditions of Corollary 2.8.3 holds independent of the channel model.

\begin{align*}
\text{Lemma 2.9.1.} & \text{ Suppose } \{\mu_n\} \subseteq \mathcal{P}(\mathbb{X}), \text{ where } \mathbb{X} \text{ is metric space, converges weakly to some } \mu \in \mathcal{P}(\mathbb{X}). \text{ For a bounded real valued sequence of functions } \{f_n\} \text{ such that } \|f_n\|_\infty < C \text{ for all } n > 0 \text{ with } C < \infty, \text{ if } \lim_{n \to \infty} f_n(x_n) = f(x) \text{ for all } x_n \to x, \text{ i.e. } f_n \text{ continuously converges to } f, \text{ then } \lim_{n \to \infty} \int_\mathbb{X} f_n(x) \mu_n(dx) = \int_\mathbb{X} f(x) \mu(dx). \\
\text{Lemma 2.9.2.} & \text{ Suppose we have a uniformly bounded family of functions } \{f^\gamma_n : \mathbb{X} \to \mathbb{R}, \gamma \in \Gamma, n > 0\} \text{ such that } \|f^\gamma_n\|_\infty < C \text{ for all } \gamma \in \Gamma \text{ and for all } n > 0 \text{ for some } C < \infty. \text{ Further suppose we have another uniformly bounded family of functions } \{f^\gamma : \mathbb{X} \to \mathbb{R}, \gamma \in \Gamma\} \text{ such that } \|f^\gamma\|_\infty < C \text{ for all } \gamma \in \Gamma \text{ for some } C < \infty. \text{ Under the following assumptions,} \hspace{1cm} (i) \text{ For any } x_n \to x \\
& \sup_{\gamma \in \Gamma} \left| f^\gamma_n(x_n) - f^\gamma(x) \right| \to 0 \quad (2.21) \\
& \sup_{\gamma \in \Gamma} \left| f^\gamma(x_n) - f^\gamma(x) \right| \to 0. \quad (2.22)
\end{align*}
(ii) $\mathcal{T}_n(\cdot|x_n, u_n)$ converges weakly to $\mathcal{T}(\cdot|x, u)$ for any $(x_n, u_n) \rightarrow (x, u)$.

(iii) $\mathcal{T}(\cdot|x, u)$ is weakly continuous in $(x, u)$.

(iv) $\mathcal{U}$ is compact.

Then for fixed observation realizations, $y_{[0,t]} := \{y_0, \ldots, y_t\}$ and for some $x^n_t \rightarrow x_t$

$$\sup_{\gamma \in \Gamma} \left| \int \mathcal{T}_n(dx_{t+1}|x^n_t, \gamma(y_{[0,t]}))f^\gamma_n(x_{t+1}) - \int \mathcal{T}(dx_{t+1}|x_t, \gamma(y_{[0,t]}))f^\gamma(x_{t+1}) \right| \rightarrow 0 \quad (2.23)$$

Proof. Using (2.22), we see that $\{f^\gamma\}$ is a equicontinuous family of functions. Thus, by the Arzela-Ascoli Theorem, for any given compact set $K \subset X$, and $\epsilon > 0$ there is a finite set of continuous functions $F := \{f^1, \ldots, f^N\}$ so that for any $\gamma$, there is $f^i \in F$ with

$$\sup_{x \in K} |f^\gamma(x) - f^i(x)| \leq \epsilon.$$ 

Now, we claim that for the same $\epsilon > 0$, the same $f^i \in F$ and the chosen compact set $K$, we can make $\sup_{x \in K} |f^\gamma_n(x) - f^i(x)| \leq 3\epsilon/2$ for large enough $n$ and for any $\gamma \in \Gamma$. To see this, observe the following:

$$\sup_{x \in K} |f^\gamma_n(x) - f^i(x)| \leq \sup_{x \in K} |f^\gamma_n(x) - f^\gamma(x)| + \sup_{x \in K} |f^\gamma(x) - f^i(x)|$$

the second term is less than $\epsilon$ with the argument we made in the first paragraph of the proof. The first term can also be made arbitrarily small since $f^\gamma_n \rightarrow f^\gamma$ uniformly on compact sets by (2.21). Now we wish to show that we can find a compact subset of $X$ such that all probability measures (kernels) in the term (2.23) put their measure mainly on this compact
set. Consider the set of measures \( S := \bigcup_{\gamma \in \Gamma} S_{\gamma} \), where

\[
S_{\gamma} = \left\{ T_n(\cdot | x^n_t, \gamma(y_{[0,t]})) : T_n(\cdot | x^n_t, \gamma(y_{[0,t]})) \rightarrow T(\cdot | x_t, \gamma(y_{[0,k]})) \right\}.
\]

Here, notice that the set \( S \) depends on the sequence \( \{x^n_t\} \) and the observation realizations \( y_{[0,t]} \). To cover all the kernels in (2.23) we take \( \{x^n_t\} \) and \( y_{[0,t]} \) as they are given in the statement of the lemma.

For a sequence from the set \( S \), say \( T_n(\cdot | x^{n_m}_t, \gamma_{m_r}(y_{[0,t]})) \); since \( U \) is a compact set and the observations are fixed, there exists a subsequence in which \( \gamma_{m_r}(y_{[0,t]}) \rightarrow u^* \) for some \( u^* \in U \). Now we focus on this subsequence \( T_{n_{m_r}}(\cdot | x^{n_{m_r}}_t, \gamma_{m_r}(y_{[0,t]})) \). By the assumption (ii) on the lemma statement \( \{T_n\} \) also satisfy the following: for any \( (x_n, u_n) \rightarrow (x, u) \), \( T_n(\cdot | x_n, u_n) \rightarrow T(\cdot | x, u) \). Thus,

\[
T_{n_{m_r}}(\cdot | x^{n_{m_r}}_t, \gamma_{m_r}(y_{[0,t]})) \rightarrow T(\cdot | x_t, u^*).
\]

Hence any sequence in the set \( S \) has a convergent subsequence, thus \( S \) is a precompact family. Therefore, it is a tight family of functions by Prokhorov theorem [14, Theorem 5.2]. Hence, for any \( \epsilon > 0 \), there exists a compact set \( K_\epsilon \) such that for all \( n \) and uniformly for all \( \gamma \in \Gamma \),

\[
\int_{K_\epsilon} T_n(dx_t | x^n_t, \gamma(y_{[0,t]})) \geq 1 - \epsilon.
\]

Now, we fix an \( \epsilon > 0 \), choose a compact set \( K_\epsilon \) according to above discussion such that all \( T_n \) puts almost all their measure (more than \( 1 - \epsilon \)) on \( K_\epsilon \). We also fix a finite family of continuous functions \( F := \{f^1, \ldots, f^N\} \) such that for any \( \gamma \), we can find an \( f^* \in F \) with

\[
\sup_{x_t \in K_\epsilon} | f^\gamma(x_t) - f^i(x_t) | \leq \epsilon.
\]

Moreover, we choose a large \( N \) such that \( \sup_{x \in K_\epsilon} | f_n^\gamma(x) - f^i(x) | \leq 3\epsilon/2 \) for all \( n \geq N \).
With this setup, we go back to the main statement:

\[
\sup_{\gamma \in \Gamma} \left| \int \mathcal{T}_n(dx_{t+1}|x^n_t, \gamma(y_{[0,t]})) f^n_i(x_{t+1}) - \int \mathcal{T}(dx_{t+1}|x_t, \gamma(y_{[0,t]})) f^\gamma(x_{t+1}) \right|
\]

\[
\leq \sup_{\gamma \in \Gamma} \left| \int_{\mathcal{X} \setminus K_e} \mathcal{T}_n(dx_{t+1}|x^n_t, \gamma(y_{[0,t]})) f^n_i(x_{t+1}) - \int_{\mathcal{X} \setminus K_e} \mathcal{T}(dx_{t+1}|x_t, \gamma(y_{[0,t]})) f^\gamma(x_{t+1}) \right|
\]

\[
+ \sup_{\gamma \in \Gamma} \left| \int_{K_e} \mathcal{T}_n(dx_{t+1}|x^n_t, \gamma(y_{[0,t]})) f^n_i(x_{t+1}) - \int_{K_e} \mathcal{T}(dx_{t+1}|x_t, \gamma(y_{[0,t]})) f^\gamma(x_{t+1}) \right|
\]

\[
\leq 2\epsilon C + \sup_{\gamma \in \Gamma} \left| \int_{K_e} \mathcal{T}_n(dx_{t+1}|x^n_t, \gamma(y_{[0,t]})) (f^n_i(x_{t+1}) - f^i(x_{t+1})) \right|
\]

\[
+ \int_{K_e} \mathcal{T}_n(dx_{t+1}|x^n_t, \gamma(y_{[0,t]})) f^i(x_{t+1}) - \int_{K_e} \mathcal{T}(dx_{t+1}|x_t, \gamma(y_{[0,t]})) f^i(x_{t+1})
\]

\[
+ \int_{K_e} \mathcal{T}(dx_{t+1}|x_t, \gamma(y_{[0,t]}))(f^i(x_{t+1}) - f^\gamma(x_{t+1})) \right|
\]

\[
\leq 2\epsilon C + \sup_{\gamma \in \Gamma} \left| \int_{K_e} \mathcal{T}_n(dx_{t+1}|x^n_t, \gamma(y_{[0,t]})) f^i(x_{t+1}) \right|
\]

\[
- \int_{K_e} \mathcal{T}(dx_{t+1}|x_t, \gamma(y_{[0,t]})) f^i(x_{t+1}) + 5\epsilon/2 \leq 4\epsilon C + 7\epsilon/2
\]

where \( C \) is the uniform bound of \( f^n_i \) and \( f^i(x_{t+1}) \) is chosen according to the discussion above such that \( f^i \) is \( \epsilon \) close to \( f^\gamma(x_{t+1}) \) and the same \( f^i \) is \( 3\epsilon/2 \) close to \( f^n_i(x_{t+1}) \).

At the last step, we used the fact that \( \mathcal{T}_n(dx_{t+1}|x^n_t, \gamma(y_{[0,t]})) \) converges weakly to \( \mathcal{T}(dx_{t+1}|x_t, \gamma(y_{[0,t]})) \) uniformly on \( \mathcal{U} \). Thus,

\[
\sup_{\gamma \in \Gamma} \left| \int_{K_e} \mathcal{T}_n(dx_{t+1}|x^n_t, \gamma(y_{[0,t]})) f^i(x_{t+1}) - \int_{K_e} \mathcal{T}(dx_{t+1}|x_t, \gamma(y_{[0,t]})) f^i(x_{t+1}) \right|
\]

\[
\leq \sup_{\gamma \in \Gamma} \left| \int_{\mathcal{X} \setminus K_e} \mathcal{T}_n(dx_{t+1}|x^n_t, \gamma(y_{[0,t]})) f^i(x_{t+1}) - \int_{\mathcal{X} \setminus K_e} \mathcal{T}(dx_{t+1}|x_t, \gamma(y_{[0,t]})) f^i(x_{t+1}) \right|
\]

\[
+ \sup_{\gamma \in \Gamma} \left| \int_{\mathcal{X}} \mathcal{T}_n(dx_{t+1}|x^n_t, \gamma(y_{[0,t]})) f^i(x_{t+1}) - \int_{\mathcal{X}} \mathcal{T}(dx_{t+1}|x_t, \gamma(y_{[0,t]})) f^i(x_{t+1}) \right| \leq 2\epsilon C + \epsilon
\]

for large enough \( n \). As \( \epsilon \) is arbitrary, the result follows.
With this lemma, we go back to (2.3). For easiness of notation we will first study the case where $t = 3$.

$$\sup_{\gamma \in \Gamma} \left| E_{p}^{T} \left[ c(X_{3}, \gamma(Y_{[0,3]})) \right] - E_{p}^{T_{n}} \left[ c(X_{3}, \gamma(Y_{[0,3]})) \right] \right|$$

$$= \sup_{\gamma \in \Gamma} \left| \int P(dx_{0})Q(dy_{0}|x_{0})T(dx_{1}|x_{0}, \gamma(y_{0}))Q(dy_{1}|x_{1})T(dx_{2}|x_{1}, \gamma(y_{[0,1]}))
\times Q(dy_{2}|x_{2})T(dx_{3}|x_{2}, \gamma(y_{[0,2]}))Q(dy_{3}|x_{3})c(x_{3}, \gamma(y_{[0,3]}))
- \int P(dx_{0})Q(dy_{0}|x_{0})T_{n}(dx_{1}|x_{0}, \gamma(y_{0}))Q(dy_{1}|x_{1})T_{n}(dx_{2}|x_{1}, \gamma(y_{[0,1]}))
\times Q(dy_{2}|x_{2})T_{n}(dx_{3}|x_{2}, \gamma(y_{[0,2]}))Q(dy_{3}|x_{3})c(x_{3}, \gamma(y_{[0,3]})) \right| .$$

Using the dominated convergence theorem, it suffices to show that

$$\sup_{\gamma \in \Gamma} \left| \int T_{n}(dx_{1}|x_{0}, \gamma(y_{0}))Q(dy_{1}|x_{1})T_{n}(dx_{2}|x_{1}, \gamma(y_{[0,1]}))
\times Q(dy_{2}|x_{2})T_{n}(dx_{3}|x_{2}, \gamma(y_{[0,2]}))Q(dy_{3}|x_{3})c(x_{3}, \gamma(y_{[0,3]}))
- \int T(dx_{1}|x_{0}, \gamma(y_{0}))Q(dy_{1}|x_{1})T(dx_{2}|x_{1}, \gamma(y_{[0,1]}))
\times Q(dy_{2}|x_{2})T(dx_{3}|x_{2}, \gamma(y_{[0,2]}))Q(dy_{3}|x_{3})c(x_{3}, \gamma(y_{[0,3]})) \right| \to 0 .$$

Then, using Lemma 2.9.2, it suffices to show that for any $x_{1}^{n} \to x_{1}$

$$\sup_{\gamma \in \Gamma} \left| \int Q(dy_{1}|x_{1}^{n})T_{n}(dx_{2}|x_{1}^{n}, \gamma(y_{[0,1]}))Q(dy_{2}|x_{2})T_{n}(dx_{3}|x_{2}, \gamma(y_{[0,2]}))Q(dy_{3}|x_{3})c(x_{3}, \gamma(y_{[0,3]}))
- \int Q(dy_{1}|x_{1})T(dx_{2}|x_{1}, \gamma(y_{[0,1]}))Q(dy_{2}|x_{2})T(dx_{3}|x_{2}, \gamma(y_{[0,2]}))Q(dy_{3}|x_{3})c(x_{3}, \gamma(y_{[0,3]})) \right| \to 0$$

(2.24)
and

\[
\sup_{\gamma \in \Gamma} \left| \int Q(dy_1|x_1^n)T(dx_2|x_1^n, \gamma(y[0,1]))Q(dy_2|x_2)T(dx_3|x_2, \gamma(y[0,2]))Q(dy_3|x_3)c(x_3, \gamma(y[0,3])) \right. \\
- \left. \int Q(dy_1|x_1)T(dx_2|x_1, \gamma(y[0,1]))Q(dy_2|x_2)T(dx_3|x_2, \gamma(y[0,2]))Q(dy_3|x_3)c(x_3, \gamma(y[0,3])) \right| \to 0. 
\]

(2.25)

We only focus on the term (2.24), the analysis for the term (2.25) follows from identical steps. For (2.24), we write the following:

\[
\sup_{\gamma \in \Gamma} \left| \int Q(dy_1|x_1^n)T_n(dx_2|x_1^n, \gamma(y[0,1]))Q(dy_2|x_2)T_n(dx_3|x_2, \gamma(y[0,2]))Q(dy_3|x_3)c(x_3, \gamma(y[0,3])) \right. \\
- \left. \int Q(dy_1|x_1)T_n(dx_2|x_1, \gamma(y[0,1]))Q(dy_2|x_2)T_n(dx_3|x_2, \gamma(y[0,2]))Q(dy_3|x_3)c(x_3, \gamma(y[0,3])) \right| \\
\leq \sup_{\gamma \in \Gamma} \left| \int Q(dy_1|x_1^n)T_n(dx_2|x_1^n, \gamma(y[0,1]))Q(dy_2|x_2)T_n(dx_3|x_2, \gamma(y[0,2]))Q(dy_3|x_3)c(x_3, \gamma(y[0,3])) \right. \\
- \left. \int Q(dy_1|x_1)T_n(dx_2|x_1^n, \gamma(y[0,1]))Q(dy_2|x_2)T_n(dx_3|x_2^n, \gamma(y[0,2]))Q(dy_3|x_3)c(x_3, \gamma(y[0,3])) \right| \\
+ \sup_{\gamma \in \Gamma} \left| \int Q(dy_1|x_1)T_n(dx_2|x_1^n, \gamma(y[0,1]))Q(dy_2|x_2)T_n(dx_3|x_2, \gamma(y[0,2]))Q(dy_3|x_3)c(x_3, \gamma(y[0,3])) \right. \\
- \left. \int Q(dy_1|x_1)T(dx_2|x_1, \gamma(y[0,1]))Q(dy_2|x_2)T(dx_3|x_2, \gamma(y[0,2]))Q(dy_3|x_3)c(x_3, \gamma(y[0,3])) \right| 
\]

(2.26)

The first term goes to 0 since the channel is continuous in total variation. For the second term, using Lemma 2.9.2 and the total variation continuity of \( Q \) successively, it reduces to show that

\[
\sup_{\gamma \in \Gamma} \left| \int Q(dy_3|x_3^n)c(x_3^n, \gamma(y[0,3])) - \int Q(dy_3|x_3)c(x_3, \gamma(y[0,3])) \right| \to 0.
\]
To show this, we write the following:

\[
\sup_{\gamma \in \Gamma} \left| \int Q(\{x_3^n\} | \{x_3\}) c(x_3^n, \gamma(y_{[0,3]})) - \int Q(\{x_3\} | \{x_3\}) c(x_3, \gamma(y_{[0,3]})) \right| \\
\leq \sup_{\gamma \in \Gamma} \left| \int Q(\{x_3^n\} | \{x_3\}) c(x_3^n, \gamma(y_{[0,3]})) - \int Q(\{x_3\} | \{x_3\}) c(x_3^n, \gamma(y_{[0,3]})) \right| \\
+ \sup_{\gamma \in \Gamma} \left| \int Q(\{x_3\} | \{x_3\}) c(x_3, \gamma(y_{[0,3]})) - c(x_3, \gamma(y_{[0,3]})) \right|.
\]

The first term goes to 0 since \(Q\) is continuous in total variation and the second term goes to 0 since \(c\) is continuous in \(x\) uniformly over \(U\). Thus, (2.24) holds true. (2.25) also holds true with identical arguments; we use the convergence of \(T_n(\cdot | x_n, u)\) to \(T(\cdot | x, u)\) uniformly over \(u \in U\) whereas for (2.24), we use the convergence of \(T_n(\cdot | x_n, u)\) to \(T(\cdot | x, u)\) uniformly over \(u \in U\) at (2.26) with Lemma 2.9.2. Therefore, (2.3) goes to 0 for the time step \(t = 3\).

For a general finite time stage \(t\), we can again use the iterative approach we used when \(t = 3\). Thus, we can generalize that for any \(0 < t < \infty\)

\[
\sup_{\gamma \in \Gamma} \left| \int T_n(dx_{t+1} | x_t, \gamma(y_{[0,t]})) f_n(x_{t+1}) - \int T(dx_{t+1} | x_t, \gamma(y_{[0,t]})) f(x_{t+1}) \right| \to 0
\]

for a fixed observation realizations \(y_{[0,t]} := \{y_0, \ldots, y_t\}\) and a fixed state \(x_t\), under the
following assumptions

(i) $T_n$ is such that for any sequence $\{u_n\} \subset U$ converging to some $u \in U$, $T_n(\cdot|u_n) \to T(\cdot|u)$ in total variation for all $x \in X$,

(ii) $T(\cdot|u)$ is continuous in total variation in $u$ for every given $x \in X$.

Proof.

\[
\sup_{\gamma \in \Gamma} \left| \int \mathcal{T}_n(dx_{t+1}|x_t, \gamma(y_{[0,t]})) f_n^\gamma(x_{t+1}) - \int \mathcal{T}(dx_{t+1}|x_t, \gamma(y_{[0,t]})) f_n^\gamma(x_{t+1}) \right| \\
\leq \sup_{\gamma \in \Gamma} \left| \int \mathcal{T}_n(dx_{t+1}|x_t, \gamma(y_{[0,t]})) f_n^\gamma(x_{t+1}) - \int \mathcal{T}(dx_{t+1}|x_t, \gamma(y_{[0,t]})) f_n^\gamma(x_{t+1}) \right| \\
+ \sup_{\gamma \in \Gamma} \left| \int \mathcal{T}(dx_{t+1}|x_t, \gamma(y_{[0,t]}))(f_n^\gamma(x_{t+1}) - f(x_{t+1})) \right| \\
\leq \sup_{\gamma \in \Gamma} \|\mathcal{T}_n(dx_{t+1}|x_t, \gamma(y_{[0,t]})) - \mathcal{T}(dx_{t+1}|x_t, \gamma(y_{[0,t]}))\|_{TV} \\
+ \sup_{\gamma \in \Gamma} \left| \int \mathcal{T}(dx_{t+1}|x_t, \gamma(y_{[0,t]}))(f_n^\gamma(x_{t+1}) - f(x_{t+1})) \right|.
\]

Above, the first term goes to 0 as $T_n(\cdot|u_n) \to T(\cdot|u)$ in total variation and $U$ is compact.

For the second term, first we use the assumption that $T(\cdot|u)$ is continuous in $u$. For any $\epsilon > 0$, there exists a $\delta > 0$ such that $|u' - u| < \delta$ implies $\|T(\cdot|u) - T(\cdot|u')\|_{TV} < \epsilon$.

Furthermore, by the assumption $U$ is compact. Therefore, for the given $\delta$, there exists a finite set $\{u_1, \cdots, u_N\}$ such that for any $\gamma \in \Gamma$, we can find a $u_i$ with $|u_i - \gamma(y_{[0,t]})| < \delta$.

Combining what we have; for any $\epsilon > 0$ and for any $\gamma \in \Gamma$, we can find a $u_i$ such that $\|T(\cdot|\gamma(y_{[0,t]})) - T(\cdot|u_i)\|_{TV} < \epsilon$. Now we focus on the second term again:

\[
\sup_{\gamma \in \Gamma} \left| \int \mathcal{T}(dx_{t+1}|x_t, \gamma(y_{[0,t]}))(f_n^\gamma(x_{t+1}) - f(x_{t+1})) \right|
\]
Thus, the first term is less than $\|c\|_\infty \epsilon$ and the second term can be made arbitrarily small for large enough $n$ since $\sup_{\gamma \in \Gamma} |f_n^\gamma(x) - f_\gamma(x)| \to 0$ by assumption. The result follows. 

Now we go back to (2.5). We will first study the case where $t = 3$.

$$
\sup_{\gamma \in \Gamma} \left| \int P(dx_0)Q(dy_0|x_0)T(dx_1|x_0, \gamma(y_0))Q(dy_1|x_1)T(dx_2|x_1, \gamma(y_{0,1}))
\times Q(dy_2|x_2)T(dx_3|x_2, \gamma(y_{0,2}))Q(dy_3|x_3)c(x_3, \gamma(y_{0,3}))
- \int P(dx_0)Q(dy_0|x_0)T_n(dx_1|x_0, \gamma(y_0))Q(dy_1|x_1)T_n(dx_2|x_1, \gamma(y_{0,1}))
\times Q(dy_2|x_2)T_n(dx_3|x_2, \gamma(y_{0,2}))Q(dy_3|x_3)c(x_3, \gamma(y_{0,3})) \right|
$$

Using Lemma 2.9.3 it suffices to show that

$$
\sup_{\gamma \in \Gamma} \left| \int \left( Q(dy_1|x_1)T_n(dx_2|x_1, \gamma(y_{0,1}))Q(dy_2|x_2)T_n(dx_3|x_2, \gamma(y_{0,2}))
- Q(dy_1|x_1)T(dx_2|x_1, \gamma(y_{0,1}))Q(dy_2|x_2)T(dx_3|x_2, \gamma(y_{0,2})) \right)
\times Q(dy_3|x_3)c(x_3, \gamma(y_{0,3})) \right| \to 0.
$$
Following the same procedure and using Lemma 2.9.3 successively, it reduces to show that

\[
\sup_{\gamma \in \Gamma} \left| \int \mathcal{T}_n(dx_3|x_2, \gamma(y_{[0,2]}))Q(dy_3|x_3)c(x_3, \gamma(y_{[0,3]})) - \int \mathcal{T}(dx_3|x_2, \gamma(y_{[0,2]}))Q(dy_3|x_3)c(x_3, \gamma(y_{[0,3]})) \right| \\
\leq \|c\|_\infty \sup_{\gamma \in \Gamma} \| \mathcal{T}_n(dx_3|x_2, \gamma(y_{[0,2]})) - \mathcal{T}(dx_3|x_2, \gamma(y_{[0,2]})) \|_{TV} \to 0
\]

which holds true by the assumptions, i.e., since the action space \( U \) is compact and \( \mathcal{T}_n \) is such that for any sequence \( \{u_n\} \subset U \) converging to some \( u \in U \), \( \mathcal{T}_n(\cdot|x, u_n) \to \mathcal{T}(\cdot|x, u) \) in total variation for all \( x \in X \). This argument can be applied to any time step \( t < \infty \).

### 2.9.3 Proof of (2.7) in Theorem 2.4.6

First, we provide the analysis for \( k = 2 \).

\[
\| P^{\gamma}_T(d(x, y, u)_{[0,2]}) - P^{\gamma}_T(d(x, y, u)_{[0,2]}) \|_{TV} \\
= \sup_{\|f\|_{\infty} \leq 1} \left| \int P(dx_0)Q(dy_0|x_0)1_{\{\gamma(0)\in du_0\}} \mathcal{T}_n(dx_1|x_0, u_0)Q(dy_1|x_1)1_{\{\gamma(y_0, y_1)\in du_1\}} \\
\times \mathcal{T}_n(dx_2|x_1, u_1)Q(dy_2|x_2)1_{\{\gamma(y_0, y_1, y_2)\in du_2\}} f(x, y, u)_{[0,2]} \right| \\
- \int P(dx_0)Q(dy_0|x_0)1_{\{\gamma(0)\in du_0\}} \mathcal{T}(dx_1|x_0, u_0)Q(dy_1|x_1)1_{\{\gamma(y_0, y_1)\in du_1\}} \\
\times \mathcal{T}(dx_2|x_1, u_1)Q(dy_2|x_2)1_{\{\gamma(y_0, y_1, y_2)\in du_2\}} f(x, y, u)_{[0,2]} \\
\leq \sup_{\|f\|_{\infty} \leq 1} \left| \int P^{\gamma}_T(d(x_{[0,1]}, y_{[0,1]}, u_{[0,1]})) \right| \int \mathcal{T}_n(dx_2|x_1, u_1)Q(dy_2|x_2)1_{\{\gamma(y_{[0,2]}\in du_2\)} f(x, y, u)_{[0,2]} \\
- \int \mathcal{T}(dx_2|x_1, u_1)Q(dy_2|x_2)1_{\{\gamma(y_{[0,2]}\in du_2\)} f(x, y, u)_{[0,2]} \\
+ \sup_{\|f\|_{\infty} \leq 1} \int P^{\gamma}_T(d(x_0, y_0, u_0)) \int \mathcal{T}_n(dx_1|x_0, u_0)Q(dy_1|x_1)1_{\{\gamma(y_0, y_1)\in du_1\}} P^{\gamma}_T(d(x_2, y_2, u_2)) f(x, y, u)_{[0,2]} \\
\]
2.9. TECHNICAL RESULTS

\[
- \int \mathcal{T}(dx_1|x_0, u_0)Q(dy_1|x_1)1_{\{y_0, u_1\} \in du_1} P_T^\gamma(d(x_2, y_2, u_2))f(x, y, u)_{[0, 2]}
\]

\[\leq 2 \sup_{x \in \mathcal{X}, u \in \mathcal{U}} \|\mathcal{T}_n(.|x, u) - \mathcal{T}(.|x, u)\|_{TV}\]

Now, we do the same analysis for a general time step \(k\):

\[
\|P_{T_n}^\gamma(d(x, y, u)_{[0, k]}) - P_T^\gamma(d(x, y, u)_{[0, k]})\|_{TV}
\]

\[= \sup_{|\mathcal{F}| \leq 1} \int f(x, y, u)_{[0, k]}P_{T_n}^\gamma(d(x, y, u)_{[0, k]}) - \int f(x, y, u)_{[0, k]}P_T^\gamma(d(x, y, u)_{[0, k]})
\]

\[\leq \sup_{|\mathcal{F}| \leq 1} \int P_{T,T_n}^\gamma(dx_0, dy_0, du_0) \left[ \int \mathcal{T}(dx_1|x_0, u_0) \int f(x, y, u)_{[0, k]}P_{T,T_n}^\gamma(dx_{[2, k]}, dy_{[1, k]}, du_{[1, k]})
\]

\[\leq \sup_{|\mathcal{F}| \leq 1} \int P_{T,T_n}^\gamma(dx_0, dy_0, du_0) \left[ \int \mathcal{T}(dx_1|x_0, u_0) \int f(x, y, u)_{[0, k]}P_{T,T_n}^\gamma(dx_{[2, k]}, dy_{[1, k]}, du_{[1, k]})
\]

\[\leq \sup_{|\mathcal{F}| \leq 1} \int P_{T,T_n}^\gamma(dx_0, dy_0, du_0) \left[ \int \mathcal{T}(dx_1|x_0, u_0) \int f(x, y, u)_{[0, k]}P_{T,T_n}^\gamma(dx_{[2, k]}, dy_{[1, k]}, du_{[1, k]})
\]

\[\leq \sup_{|\mathcal{F}| \leq 1} \int P_{T,T_n}^\gamma(dx_0, dy_0, du_0) \left[ \int \mathcal{T}(dx_1|x_0, u_0) \int f(x, y, u)_{[0, k]}P_{T,T_n}^\gamma(dx_{[2, k]}, dy_{[1, k]}, du_{[1, k]})
\]

\[\leq k \sup_{x \in \mathcal{X}, u \in \mathcal{U}} \|\mathcal{T}(.|x, u) - \mathcal{T}_n(.|x, u)\|_{TV}.
\]

In the argument above \(P_{T,T_n}^\gamma\) denotes a strategic measure that uses \(\mathcal{T}\) or \(\mathcal{T}_n\) at various steps.

The terms are arranged so that at every term the applied strategic measures coincide.
2.9. TECHNICAL RESULTS

2.9.4 Proof for Remark 2.6.3

We note that if the family where we search for policies is restricted to an equicontinuous family of functions, robustness can also be achieved. Let $\Gamma_{eq}$ be the family of equicontinuous policies so that for any given $x_0 \in X$ and $\epsilon > 0$, there exists a $\delta > 0$ such that $|\gamma(x) - \gamma(x_0)| \leq \epsilon$ for all $\gamma \in \Gamma_{eq}$ and for every $x$ such that $|x - x_0| \leq \delta$.

We show that for all $t < \infty$

$$\sup_{\gamma \in \Gamma_{eq}} \left| E^T \left[ c(X_t, \gamma(X_t)) \right] - E^T_n \left[ c(X_t, \gamma(X_t)) \right] \right| \to 0.$$  

For ease of notation we will first study the case where $t = 2$.

$$\sup_{\gamma \in \Gamma} \left| E^T \left[ c(X_2, \gamma(X_2)) \right] - E^T_n \left[ c(X_2, \gamma(X_2)) \right] \right|$$

$$= \sup_{\gamma \in \Gamma} \left| \int \mathcal{T}(dx_1|x_0, \gamma(x_0))\mathcal{T}(dx_2|x_1, \gamma(x_1))c(x, \gamma(x_2)) \right.$$ 

$$- \left. \int \mathcal{T}_n(dx_1|x_0, \gamma(x_0))\mathcal{T}_n(dx_2|x_1, \gamma(x_1))c(x_2, \gamma(x_2)) \right|.$$ 

To show that above term goes to 0, we use a lemma.

**Lemma 2.9.4.** Suppose we have a uniformly bounded family of functions $\{f_n^\gamma : X \to \mathbb{R}, \gamma \in \Gamma_{eq}, n > 0\}$ such that $\|f_n^\gamma\|_\infty < C$ for all $\gamma \in \Gamma_{eq}$ and for all $n > 0$ for some $C < \infty$.

Further suppose we have another uniformly bounded family of functions $\{f^\gamma : X \to \mathbb{R}, \gamma \in \Gamma_{eq}\}$ such that $\|f^\gamma\|_\infty < C$ for all $\gamma \in \Gamma_{eq}$ for some $C < \infty$. Under the following assumptions,
(i) For any $x_n \to x$

\[ \sup_{\gamma \in \Gamma_{eq}} \left| f_n^\gamma(x_n) - f^\gamma(x) \right| \to 0 \]  \hspace{1cm} (2.27)

\[ \sup_{\gamma \in \Gamma_{eq}} \left| f(x) - f^\gamma(x) \right| \to 0. \]  \hspace{1cm} (2.28)

(ii) $T_n(\cdot|x_n, u_n)$ converges weakly to $T(\cdot|x, u)$ for any $(x_n, u_n) \to (x, u)$.

(iii) $T(\cdot|x, u)$ is weakly continuous in $(x, u)$.

(iv) $U$ is compact.

Then for some $x^n_t \to x_t$

\[ \sup_{\gamma \in \Gamma_{eq}} \left| \int T_n(dx_{t+1}|x^n_t, \gamma(x^n_t)) f_n^\gamma(x^n_t) - \int T(dx_{t+1}|x_t, \gamma(x_t)) f^\gamma(x_{t+1}) \right| \to 0. \]  \hspace{1cm} (2.29)

**Proof.** Using the same steps as in Lemma A.2 we can show that for any given compact set $K \subset X$, and $\epsilon > 0$ there is a finite set of continuous functions $F := \{f^1, \ldots, f^N\}$ so that for any $\gamma$, there is $f^i \in F$ with

\[ \sup_{x \in K} |f^\gamma(x) - f^i(x)| \leq \epsilon. \]

For the same $\epsilon > 0$, the same $f^i \in F$ and the chosen compact set $K$, we can also make

\[ \sup_{x \in K} \left| f_n^\gamma(x) - f^i(x) \right| \leq 3\epsilon/2 \text{ for large enough } n \text{ and for any } \gamma \in \Gamma_{eq}. \]

We also that the set of measures $S := \bigcup_{\gamma \in \Gamma_{eq}} S_\gamma$ is weakly compact where

\[ S_\gamma = \{ T_n(\cdot|x^n_t, \gamma(x^n_t)) : T_n(\cdot|x^n_t, \gamma(x^n_t)) \to T(\cdot|x_t, \gamma(x_t)) \}. \]
Hence, for any $\epsilon > 0$, there exists a compact set $K_\epsilon$ such that for all $n$ and uniformly for all $\gamma \in \Gamma,$
\[
\int_{K_\epsilon} \mathcal{T}_n(dx_t^n|x_t^n, \gamma(x_t^n)) \geq 1 - \epsilon.
\]

Using these, we again follow the same steps as in the proof of Lemma A.2:

\[
\begin{align*}
\sup_{\gamma \in \Gamma_{eq}} \left| \int \mathcal{T}_n(dx_{t+1}^n|x_t^n, \gamma(x_t^n))f_n^n(x_{t+1}) - \int \mathcal{T}(dx_{t+1}|x_t, \gamma(x_t))f^\gamma(x_{t+1}) \right| \\
\leq \sup_{\gamma \in \Gamma_{eq}} \left| \int_{\mathcal{X}\setminus K_\epsilon} \mathcal{T}_n(dx_{t+1}^n|x_t^n, \gamma(x_t^n))f_n^n(x_{t+1}) - \int_{\mathcal{X}\setminus K_\epsilon} \mathcal{T}(dx_{t+1}|x_t, \gamma(x_t))f^\gamma(x_{t+1}) \right| \\
+ \sup_{\gamma \in \Gamma_{eq}} \left| \int_{K_\epsilon} \mathcal{T}_n(dx_{t+1}^n|x_t^n, \gamma(x_t^n))f_n^n(x_{t+1}) - \int_{K_\epsilon} \mathcal{T}(dx_{t+1}|x_t, \gamma(x_t))f^\gamma(x_{t+1}) \right| \\
\leq 2\epsilon C + \sup_{\gamma \in \Gamma_{eq}} \left| \int_{K_\epsilon} \mathcal{T}_n(dx_{t+1}^n|x_t^n, \gamma(x_t^n))(f_n^n(x_{t+1}) - f^i(x_{t+1})) \right| \\
+ \int_{K_\epsilon} \mathcal{T}_n(dx_{t+1}^n|x_t^n, \gamma(x_t^n))f^i(x_{t+1}) - \int_{K_\epsilon} \mathcal{T}(dx_{t+1}|x_t, \gamma(x_t))f^i(x_{t+1}) \\
+ \int_{K_\epsilon} \mathcal{T}(dx_{t+1}|x_t, \gamma(x_t))(f^i(x_{t+1}) - f^\gamma(x_{t+1})) \right| \\
\leq 2\epsilon C + \sup_{\gamma \in \Gamma_{eq}} \left| \int_{K_\epsilon} \mathcal{T}_n(dx_{t+1}^n|x_t^n, \gamma(x_t^n))f^i(x_{t+1}) \right| \\
- \int_{K_\epsilon} \mathcal{T}(dx_{t+1}|x_t, \gamma(x_t))f^i(x_{t+1}) \right| + 5\epsilon/2 \leq 4\epsilon C + 7\epsilon/2
\end{align*}
\]

where $C$ is the uniform bound of $f_n^n$ and $f^i(x_{t+1})$ is chosen according to the discussion above such that $f^i$ is $\epsilon$ close to $f^\gamma(x_{t+1})$ and the same $f^i$ is $3\epsilon/2$ close to $f_n^n(x_{t+1}).$

At the last step, we used the fact that $\mathcal{T}_n(dx_{t+1}^n|x_t^n, \gamma(x_t^n))$ converges weakly to $\mathcal{T}(dx_{t+1}|x_t, \gamma(x_t))$ uniformly over $\Gamma_{eq}$ as $\Gamma_{eq}$ is equicontinuous. As $\epsilon$ is arbitrary, the result follows. \qed

Now we go back to
\[
\sup_{\gamma \in \Gamma} \left| E^T \left[ c(X_2, \gamma(X_2)) \right] - E^{\mathcal{T}_n} \left[ c(X_2, \gamma(X_2)) \right] \right|
\]
\[ = \sup_{\gamma \in \Gamma} \left| \int \mathcal{T}(dx_1|x_0, \gamma(x_0))\mathcal{T}(dx_2|x_1, \gamma(x_1))c(x_2, \gamma(x_2)) \\
- \int \mathcal{T}_n(dx_1|x_0, \gamma(x_0))\mathcal{T}_n(dx_2|x_1, \gamma(x_1))c(x_2, \gamma(x_2)) \right|. \]

The previous lemma can be used to show that this term converges to 0.
Chapter 3

Continuity and Robustness under Incorrect Transition Kernels under Infinite Horizon Average Cost Criterion

In this chapter, we study the same problem we have studied in Chapter 2 under the infinite horizon average cost criterion. That is, we study continuity properties of discrete-time stochastic control problems with respect to system models (i.e., controlled transition kernels) and robustness of optimal control policies designed for incorrect models applied to the true system under infinite horizon average cost criterion. In this chapter, we only focus on the fully observed settings.

The average infinite horizon cost function is defined as:

$$J_\infty(P, \mathcal{T}, \gamma) = \limsup_{N \to \infty} \frac{1}{N} \sum_{t=0}^{N-1} E_{P}^{T,\gamma} [c(X_t, U_t)],$$

over the set of admissible policies $\gamma \in \Gamma$, where $c : X \times U \to \mathbb{R}$ is a Borel-measurable stage-wise cost function and $E_{P}^{T,\gamma}$ denotes the expectation with initial state probability measure $P$ and transition kernel $\mathcal{T}$ under policy $\gamma$. 
The optimal cost as a function of the priors and the transition kernels is defined as

$$J^*_\infty(P, \mathcal{T}) = \inf_{\gamma \in \Gamma} J_\infty(P, \mathcal{T}, \gamma).$$

The restate the problems of interest:

**Problem P1: Continuity of** $J^*_\infty(P, \mathcal{T})$ **under the convergence of the transition kernels.**

Let $\{\mathcal{T}_n, n \in \mathbb{N}\}$ be a sequence of transition kernels which converge in some sense to another transition kernel $\mathcal{U}$. Does that imply that

$$J^*_\infty(P, \mathcal{U}_n) \rightarrow J^*_\infty(P, \mathcal{U})?$$

**Problem P2: Robustness to incorrect models.** Suppose that an optimal policy is constructed according to a model which is incorrect: how does the application of the control to the true model affect the system performance and does the error decrease to zero as the models become closer to each other? In particular, suppose that $\gamma^*_n$ is an optimal policy designed for $\mathcal{T}_n$, an incorrect model for a true model $\mathcal{T}$. Is it the case that if $\mathcal{T}_n \rightarrow \mathcal{T}$ then

$$J_\infty(P, \mathcal{T}, \gamma^*_n) \rightarrow J^*_\infty(P, \mathcal{T})?$$

### 3.1 Literature Review and Contributions

As we have noted in Chapter 2, robustness in controlled systems is a classical problem, so there is a very large literature on robust stochastic control and its application to learning-theoretic methods. The reader can refer to Chapter 2 for a detailed review of related literature; here we list some of them for convenience see e.g. [58, 34, 83, 86, 37, 36, 102, 66, 105].

A particularly relevant study which has investigated the robustness problem under the
expected average-cost criterion is \cite{51} by Hernandez-Lerma: In \cite{51} Chapter 3, a related problem to what we study in this chapter is considered: Assume the transition probabilities are estimated through time, and at every time step $t$ the estimate is updated to some $T_t$. Is it true that, as the estimates through time get closer to the true model (as $T_t \to T$) we can establish continuity and robustness? It is shown in \cite{51} Chapter 3] that, using some geometric ergodicity conditions and with uniform convergence \( \sup_{x,u} \| T_n(\cdot | x, u) - T(\cdot | x, u) \|_{TV} \to 0 \), the answer to the question is positive. In Section 3.5, we will show that the conditions imposed in \cite{51} Chapter 3 are more restrictive than what we will present in this chapter.

**Contributions.** In view of the literature review reported above, we establish robustness and continuity results under much weaker conditions on the proximity and convergence properties between incorrect models and a true model: we show that the average optimal cost is continuous in the convergences of controlled transition kernel models where convergences of models entail (i) continuous weak convergence in state and actions, or (ii) continuous setwise convergence in the actions for every fixed state variable, in addition to ergodicity conditions. We show that the mismatch error due to the application of a control policy designed for an incorrectly estimated model to the true model decreases to zero as the incorrect model approaches the true model under the stated convergence criteria (Theorem 3.4.2, Theorem 3.4.3 and Theorem 3.4.4). We will also establish an ergodic invariant measure argument, in Theorem 3.6.1 for cases where the average cost optimality equation cannot be established. These findings are applied to empirical consistency results in data-driven stochastic control, where continuous weak convergence will be shown to be particularly relevant building on a measure concentration analysis. Compared to the results in \cite{51} by Hernandez-Lerma, where a uniform total variation convergence over state and
control action variables of the transition kernels is provided as a sufficient condition, we show that the uniform convergence on the state variable may be relaxed under the total variation convergence, and more importantly, we establish continuity and robustness under weak convergence and setwise convergences of the transition kernels which are much more relaxed notions of convergence.

3.2 Convergence Criteria for Transition Kernels

We use the same convergence notions that we have used in Chapter 2. Recall Definition 2.2.1:

**Definition 3.2.1.** For a sequence of transition kernels \( \{T_n, n \in \mathbb{N}\} \), we say that

(i) \( T_n \to T \) weakly if \( T_n(\cdot|x,u) \to T(\cdot|x,u) \) weakly, for all \( x \in X \) and \( u \in U \).

(ii) \( T_n \to T \) setwise if \( T_n(\cdot|x,u) \to T(\cdot|x,u) \) setwise, for all \( x \in X \) and \( u \in U \).

(iii) \( T_n \to T \) under the total variation distance if \( T_n(\cdot|x,u) \to T(\cdot|x,u) \) under total variation, for all \( x \in X \) and \( u \in U \).

We refer the reader to Example 2.2.2 for various examples for which these convergence notions are satisfied.

3.3 Supporting Results: Continuity under Convergence of Transition Kernels

3.3.1 Some Differences with the Infinite Horizon Discounted Problem

In Chapter 2, we studied continuity of infinite horizon discounted cost problem under the convergence of transition kernels. In the following, we first show that the sufficient conditions presented in Chapter 2 to guarantee the continuity may not be sufficient for the infinite
3.3. SUPPORTING RESULTS: CONTINUITY UNDER CONVERGENCE OF TRANSITION KERNELS

horizon average cost setup.

We restate Theorem 2.6.2 from Section that gives sufficient conditions to guarantee the continuity of optimal discounted cost function under weak convergence of transition kernels.

**Theorem 3.3.1.** Assume the following assumptions hold:

(i) $T_n(\cdot| x_n, u_n) \to T(\cdot| x, u)$ weakly for every $x \in X$ and $u \in U$ and $(x_n, u_n) \to (x, u)$,

(ii) $T(\cdot| x, u)$ is weakly continuous in $(x, u)$,

(iii) The cost function $c(x, u)$ is continuous and bounded in $X \times U$,

(iv) The action space $U$ is compact.

Then $J_{\beta}(T_n, \gamma^*_n) \to J_{\beta}(T, \gamma^*)$, for any initial state $x_0$, as $n \to \infty$.

In the following, we show that these conditions in Theorem 3.3.1 may not be sufficient for continuity for average cost problems. The following example shows that even for a control-free setup even if the assumptions of the above theorem hold, infinite horizon average cost function is not continuous. This example also covers the controlled case via using trivial control.

**Example 3.3.1.** Assume that $x_0 = 0$ and the transition kernels are given by

$$T(\cdot| x) = \delta_x(\cdot), \quad T_n(\cdot| x) = \delta_{x + \frac{1}{n}}(\cdot).$$

The cost function is given as

$$c(x) = \begin{cases} |x| & \text{if } |x| \leq 1, \\ 1 & \text{if } |x| > 1. \end{cases}$$
3.3. SUPPORTING RESULTS: CONTINUITY UNDER CONVERGENCE OF TRANSITION KERNELS

Notice that $\mathcal{T}_n(\cdot|x_n) \to \mathcal{T}(\cdot|x)$ weakly for any $x_n \to x$ and $\mathcal{T}(\cdot|x)$ is weakly continuous. It is easy to see that the cost for $\mathcal{T}$ is 0 as the state always stays at 0 that is $J_\infty(\mathcal{T}) = 0$. The cost for $\mathcal{T}_n$ can be calculated as follows:

$$J_\infty(\mathcal{T}_n) = \lim_{N \to \infty} \frac{1}{N} \left( \sum_{k=0}^{n} \frac{k}{n} + \sum_{k=n+1}^{N} 1 \right)$$

$$= \lim_{N \to \infty} \frac{1}{N} \left( \frac{n + 1}{2} + N - n - 1 \right) = 1 \neq 0.$$

3.3.2 Ergodicity Properties of Controlled Markov Chains

In the following we will denote the set of all stationary policies by $\Gamma_s$. For the transitions under some stationary policy $\gamma$, we will use the following notation: $\mathcal{T}(\cdot|x,\gamma) := \mathcal{T}(\cdot|x,\gamma(x))$.

We also define the $t$-step transition kernel $\mathcal{T}^t(\cdot|x,\gamma)$ in an iterative fashion as follows:

$$\mathcal{T}^t(\cdot|x,\gamma) := \int \mathcal{T}(\cdot|x_{t-1},\gamma) \mathcal{T}^{t-1}(dx_{t-1}|x,\gamma),$$

where $\mathcal{T}^1(\cdot|x,\gamma) = \mathcal{T}(\cdot|x,\gamma)$.

We will use the following ergodicity condition for some of our results.

**Assumption 3.3.1.** For every stationary policy $\gamma$, the transition kernels $\mathcal{T}$ and $\mathcal{T}_n$ lead to positive Harris recurrent chains and in particular admit invariant measures $\pi_\gamma$ and $\pi_n^\gamma$, and for these invariant measures uniformly for every initial point $x \in \mathcal{X}$ we have:

$$\lim_{t \to \infty} \sup_{\gamma \in \Gamma_s} \|\mathcal{T}^t(\cdot|x,\gamma) - \pi_\gamma(\cdot)\|_{TV} = 0$$
\[
\lim_{t \to \infty} \sup_n \sup_{\gamma \in \Gamma_x} \| T^t_n(x, \gamma) - \pi^n_\gamma(\cdot) \|_{TV} = 0.
\]

For the results in this section, we will make use of Assumption 3.3.1. However, using [54, Theorem 3.2], presented in the technical results section as Theorem 3.8.1, alternative conditions can also be used. Notice that condition Theorem 3.8.1(i) is the same as Assumption 3.3.1, thus one can make use of different assumptions through the relations provided in Theorem 3.8.1.

### 3.3.3 Optimality of Stationary Policies

For our continuity and robustness results, it will be instrumental to work with stationary policies. This will be without any loss under mild conditions to be presented in this subsection. An approach for average cost problems is to make use of average cost optimality equation (ACOE). To work with ACOE one usually needs contraction properties of the transition kernel. The following result provides further alternative sufficient conditions on existence of optimal policies (which turn out to be stationary) for infinite horizon average cost problems.

**Assumption 3.3.2.**

(A) The condition \( f \) in Theorem 3.8.1 (or any other suitable condition among a-i) holds,

(B) The action space \( \mathbb{U} \) is compact,

(C) \( c(x, u) \) is bounded and continuous in \( (x, u) \),

(C\text{'}\) \( c(x, u) \) is bounded and continuous in \( u \) for every fixed \( x \),

(D) \( T(\cdot| x, u) \) is weakly continuous in \( (x, u) \),
(D’) $T(\cdot | x, u)$ is setwise continuous in $u$ for every $x$.

**Theorem 3.3.2.** [51, Corollary 3.6] Suppose Assumption 3.3.2 A, B, and, either C and D, or $C’$ and $D’$, hold. Then $J_\infty(T, \gamma)$ admits an optimal stationary policy.

For the rest of the section, we will assume that the optimal policies can be selected from those which are stationary with suitable assumptions and we will denote the family of stationary polices by $\Gamma$.

### 3.3.4 Approximation by Finite Horizon Cost

We denote the $t$-step finite horizon cost function under a stationary policy $\gamma$ and a transition model $T$ by $J_t(T, \gamma)$ and the corresponding optimal cost is denoted by $J^*_t(T)$:

$$J_t(T, \gamma) = \sum_{i=0}^{t-1} E_{\gamma}^T [c(X_i, U_i)]$$

$$J^*_t(T) = \inf_{\gamma \in \Gamma} J_t(T, \gamma).$$

The following result shows that the infinite horizon average cost induced by a stationary policy can be approximated by a finite cost under the same stationary policies with proper ergodicity conditions.

**Lemma 3.3.1.** Under Assumption 3.3.1 if the cost function $c$ is bounded then for every initial state we have

$$\sup_{\gamma \in \Gamma} \left| \frac{J_t(T, \gamma)}{t} - J_\infty(T, \gamma) \right| \to 0,$$

$$\sup_{\gamma \in \Gamma} \sup_n \left| \frac{J_t(T_n, \gamma)}{t} - J_\infty(T_n, \gamma) \right| \to 0.$$
3.3. SUPPORTING RESULTS: CONTINUITY UNDER CONVERGENCE OF TRANSITION KERNELS

**Proof.** We have that \( J_\infty(\mathcal{T}, \gamma) = \int c(x, \gamma(x))\pi^\gamma(dx) \). Thus, we can write

\[
\left| \frac{J_t(\mathcal{T}, \gamma)}{t} - J_\infty(\mathcal{T}, \gamma) \right| = \left| \frac{1}{t} \sum_{i=0}^{t-1} E_i^T[c(X_i, U_i)] - \int c(x, \gamma(x))\pi^\gamma(dx) \right|
\]

\[
\leq \frac{1}{t} \sum_{i=0}^{t-1} \left| \int c(x_i, \gamma(x_i))\mathcal{T}^i(dx_i|x_0, \gamma) - \int c(x, \gamma(x))\pi^\gamma(dx) \right|
\]

\[
\leq \frac{1}{t} \sum_{i=0}^{t-1} \|c\|_\infty \|\mathcal{T}^i(\cdot|x_0, \gamma) - \pi^\gamma\|_{TV}.
\]

We now fix an \( \epsilon > 0 \) and choose a \( t_\epsilon < \infty \) such that \( \|\mathcal{T}^i(\cdot|x_0, \gamma) - \pi^\gamma\|_{TV} < \epsilon \) for all \( i > t_\epsilon \). We also choose another \( T_\epsilon \) with \( \frac{2t_\epsilon}{t} < \epsilon \) for all \( t > T_\epsilon \). With this setup, we have

\[
\frac{1}{t} \sum_{i=0}^{t-1} \|\mathcal{T}^i(\cdot|x_0, \gamma) - \pi^\gamma\|_{TV} \leq \frac{1}{t} \sum_{i=0}^{t_\epsilon-1} \|\mathcal{T}_\gamma^i(\cdot|x_0) - \pi^\gamma\|_{TV} + \frac{1}{t} \sum_{i=t_\epsilon}^{t} \|\mathcal{T}_\gamma^i(\cdot|x_0) - \pi^\gamma\|_{TV}
\]

\[
\leq \frac{2t_\epsilon}{t} + \epsilon \leq 2\epsilon, \quad \forall t > T_\epsilon.
\]

We have shown that for any fixed \( \epsilon > 0 \), we can choose a \( T_\epsilon < \infty \), independent of \( \gamma \), such that

\[
\left| \frac{J_t(\mathcal{T}, \gamma)}{t} - J_\infty(\mathcal{T}, \gamma) \right| < \epsilon, \quad \forall t > T_\epsilon.
\]

Hence the result is complete for \( \mathcal{T} \).

For \( \mathcal{T}_n \), the result follows from the same steps since we can again choose such \( t_\epsilon \) and \( T_\epsilon \) due to the uniformity over \( n \) and \( \gamma \) in Assumption [3.3.1].

The next result from [51, Corollary 4.11] shows that the optimal infinite horizon cost can be approximated by an optimal finite horizon cost induced by the same transition kernel.
3.3. SUPPORTING RESULTS: CONTINUITY UNDER CONVERGENCE OF TRANSITION KERNELS

Lemma 3.3.2. Suppose the cost function \( c \) is bounded and either Assumption 3.3.2 A, B, C, D or Assumption 3.3.2 A, B, C', D' hold (for \( T \) and \( T_n \)). Then, we have

\[
\lim_{t \to \infty} \left| \frac{J^*_\infty(T)}{t} - J^*_t(T) \right| = 0,
\]

\[
\lim_{t \to \infty} \sup_{n} \left| \frac{J^*_\infty(T_n)}{t} - J^*_t(T_n) \right| = 0.
\]

3.3.5 Continuity under the Convergence of Transition Kernels

Theorem 3.3.3. We have that \( |J^*_\infty(T_n) - J^*_\infty(T)| \to 0 \), under

1. Assumption 3.3.2 A, B, C and D if \( T_n(\cdot|x_n, u_n) \to T(\cdot|x, u) \) weakly for any \((x_n, u_n) \to (x, u)\).

2. Assumption 3.3.2 A, B, C' and D' if \( T_n(\cdot|x, u_n) \to T(\cdot|x, u) \) setwise for any \( u_n \to u \) for every fixed \( x \).

Proof. We use the following bound:

\[
|J^*_\infty(T_n) - J^*_\infty(T)| \leq \left| J^*_\infty(T_n) - \frac{J^*_t(T_n)}{t} \right| + \left| \frac{J^*_t(T_n)}{t} - \frac{J^*_t(T)}{t} \right| + \left| \frac{J^*_t(T)}{t} - J^*_\infty(T) \right|.
\]

The first and the last terms above can be made arbitrarily small by choosing \( t \) large enough uniformly over \( n \) using Lemma 3.3.2 under suitable assumptions. For the second term, we can use continuity results for finite time problems for the fixed \( t \) as the assumptions cover the requirements of ([64, Theorem 4.2] or [64, Section 5.3]).

Remark 3.3.1. We note that any condition set provided for continuity of the optimal cost function under setwise convergence of the transition kernels is also a sufficient set of conditions for the continuity of the optimal cost function under total variation convergence of
3.4 ROBUSTNESS TO INCORRECT CONTROLLED TRANSITION KERNEL MODELS

In this section, we investigate robustness for infinite horizon average cost problems. We first restate the problem: Consider a MDP with transition kernel $T_n$, and assume that an optimal control policy for this MDP under the average cost criterion is $\gamma^*_n$, that is

$$\inf_{\gamma \in \Gamma} J_{\infty}(T_n, \gamma) = J_{\infty}(T_n, \gamma^*_n).$$

Now, consider another MDP with transition kernel $T$ whose the optimal cost denoted by $J^*_{\infty}(T)$. The question we ask is the following: if the controller does not know the true transition kernel $T$ and calculates an optimal policy assuming the transition kernel is $T_n$, then the incurred cost by this policy is $J_{\infty}(T, \gamma^*_n)$. The focus of this section is to find sufficient conditions such that as $T_n \rightarrow T$,

$$J_{\infty}(T, \gamma^*_n) \rightarrow J_{\infty}(T, \gamma^*).$$

The first issue with this question is the following one: assume that the MDP with kernel $T_n$ admits two different optimal policies $\gamma^1_n$ and $\gamma^2_n$. Although, the cost incurred by these policies under the kernel $T_n$ are the same, under the kernel $T$ they may have different cost values. That is, even though we have that

$$J_{\infty}(T_n, \gamma^1_n) = J_{\infty}(T_n, \gamma^2_n) = J^*_{\infty}(T_n),$$
we may have $J_\infty(\mathcal{T}, \gamma_n^1) \neq J_\infty(\mathcal{T}, \gamma_n^2)$. An example is as follows: Consider a system with state space $\mathbb{X} = [-1, 1]$, control action space $\mathbb{U} = \{-1, 0, 1\}$, the cost function $c(x, u) = (x - u)^2$ and the transition models given as

$$
\mathcal{T}_n(\cdot|x, u) = \frac{1}{2} \delta_1(\cdot) + \frac{1}{2} \delta_{-1}(\cdot)
$$

$$
\mathcal{T}(\cdot|x, u) = \delta_0(\cdot)
$$

Notice that two optimal policies for $\mathcal{T}_n$ are

$$
\gamma_n^1(x) = \begin{cases} 
1 & \text{if } x = 1, \\
-1 & \text{if } x = -1, \\
0 & \text{else.}
\end{cases}
\quad \text{and} \quad
\gamma_n^2(x) = \begin{cases} 
1 & \text{if } x \geq 0, \\
-1 & \text{if } x < 0.
\end{cases}
$$

However, if the initial point is $x_0 = 0$, we have that $J_\infty(\mathcal{T}, \gamma_n^1) = 0 \neq 1 = J_\infty(\mathcal{T}, \gamma_n^2)$.

In what follows, we show that under total variation convergence of $\mathcal{T}_n \rightarrow \mathcal{T}$, this issue does not cause a problem so that we have $J_\infty(\mathcal{T}, \gamma_n^*) \rightarrow J_\infty(\mathcal{T}, \gamma^*)$ for any stationary optimal policy $\gamma_n^*$. However, under weak convergence of the transition models, we establish the same result under some particularly constructed optimal polices $\gamma_n^*$, namely we focus on the policies that solve the average cost optimality equation (ACOE).

### 3.4.1 Robustness under Weak Convergence of Transition Kernels

**The Average Cost Optimality Equation (ACOE)**

Now, we discuss the average cost optimality equation, and we use it for analyzing robustness properties of MDPs under weak convergence of the transition kernels. Define the
operator $T : B(X) \to B(X)$ where $B(X)$ denotes the set of bounded and measurable functions on $X$ such that for $v \in B(X)$

$$Tv(x) := \inf_{u \in U} \left( c(x,u) + \int_X v(y)\mathcal{T}(dy|x,u) \right).$$  \hspace{1cm} (3.1)

We define the span semi-norm of a function $v \in B(X)$ by

$$sp(v) := \sup_x v(x) - \inf_x v(x).$$

One can show that the operator defined in (3.1) is a contraction in $B(X)$ under the span-norm with Assumption 3.3.1 or any suitable one from Assumption 3.8.1 [51, Lemma 3.5]. Hence, according to the Banach fixed point theorem there exists a fixed point $v^* \in B(X)$ such that $sp(Tv^* - v^*) = 0$. By the definition of the span-norm, $Tv^*(x) - v^*(x) = j^*$ for a constant $j^*$ for all $x \in X$. That is

$$j^* + v^*(x) = \inf_{u \in U} \left( c(x,u) + \int_X v^*(y)\mathcal{T}(dy|x,u) \right).$$ \hspace{1cm} (3.2)

This constant $j^*$ is the optimal infinite horizon average cost, and equation (3.2) is called the average cost optimality equation (ACOE). For the remainder of this section, we will sometimes use the notation $J_\infty(\mathcal{T}, \gamma, x)$ or $J_{\infty}^*(\mathcal{T}, x)$ for the expected average cost or for the expected optimal cost when the process starts from the initial state $x$ in order to emphasize the initial state.

Now, we formalize these observations:

**Theorem 3.4.1.** [51] Suppose the cost function $c$ is bounded. Under Assumption 3.3.1 there exists a $\beta < 1$ such that the following holds:
3.4. ROBUSTNESS TO INCORRECT CONTROLLED TRANSITION KERNEL MODELS

(i) \( sp(Tv - Tw) \leq \beta sp(u - w) \), for any \( v, w \in B(X) \) where \( T \) is the operator defined in (3.7).

(ii) Since \( T \) is a contraction under the span norm, it admits a fixed point \( v^* \in B(X) \) such that

\[
j^* + v^*(x) = \inf_{u \in U} \left( c(x, u) + \int_X v^*(y)T(dy|x, u) \right),
\]

for some constant \( j^* \).

(iii) For any initial point \( x_0 \in X \), the constant \( j^* \) defined in (ii) is the optimal infinite horizon average cost for the kernel \( T \), that is

\[
j^* = J^*_\infty(T, x_0) = \inf_{\gamma \in \Gamma} J^*_\infty(T; \gamma, x_0)
\]

for every \( x_0 \in X \).

(iv) If there exists a policy \( \gamma^* \in \Gamma \) satisfying the ACOE, then this stationary policy is an optimal policy for the average infinite horizon cost problem; that is, if \( \gamma^* \) satisfies

\[
j^* + v^*(x) = c(x, \gamma^*(x)) + \int_X v^*(y)T(dy|x, \gamma^*(x)),
\]

then \( J^*_\infty(T, \gamma^*, x_0) = J^*_\infty(T, x_0) \).

We now state the main result of this section.

**Theorem 3.4.2.** We have that

\[
J^*_\infty(T, \gamma^*_n, x) \to J^*_\infty(T, x)
\]
for any \( x \in X \), where \( \gamma^*_n \) is the optimal policy for the transition kernel \( T_n \) that satisfies the ACOE, under Assumption 3.3.2 A, B, C and D if \( T_n(\cdot|x_n, u_n) \to T(\cdot|x, u) \) weakly for any \((x_n, u_n) \to (x, u)\).

**Proof.** Consider the following two ACOEs for the kernels \( T_n \) and \( T \) with their fixed points \( v^*_n \) and \( v^* \):

\[
\begin{align*}
    j^*_n + v^*_n(x) &= \inf_{u \in U} \left[ c(x, u) + \int v^*_n(y) T_n(dy|x, u) \right] \quad (3.3) \\
    j^* + v^*(x) &= \inf_{u \in U} \left[ c(x, u) + \int v^*(y) T(dy|x, u) \right] \quad (3.4)
\end{align*}
\]

We now show that, for all \( x_n \to x \),

\[
v^*_n(x_n) - v^*(x) \to c \quad (3.5)
\]

for some constant \( c \) with \(|c| < \infty\). To show this, we first write

\[
v^*_n(x_n) - v^*(x) = (v^*_n(x_n) - v^t_n(x_n)) + (v^t_n(x_n) - v^t(x)) + (v^t(x) - v^*(x))
\]

where \( v^t_n \) and \( v^t \) are the results of operator (3.1) applied to the 0-function, \( t \) times for kernels the \( T_n \) and \( T \). Notice that \( v^t_n \) and \( v^t \) are the value functions for \( t \)-step cost problem and by the assumptions ([64, Theorem 4.4]) we have that \(|v^t_n(x_n) - v^t(x)| \to 0\) for every fixed \( t \). For the first and the last terms, we use the fact that the operator (3.1) is a contraction under Assumption 3.8.1 for span semi-norm and hence both terms go to some constants as \( t \to \infty \) uniformly for all \( n \), that is \( v^*_n(x_n) - v^t_n(x_n) \to c_1 \) and \( v^t(x) - v^*(x) \to c_2 \) for some \(|c_1|, |c_2| < \infty\). Thus, we have that (3.5) holds for some \( c < \infty \).
3.4. ROBUSTNESS TO INCORRECT CONTROLLED TRANSITION KERNEL MODELS

Since \( U \) is compact, for every \( x_n \to x \), \( \gamma_n(x_n) \) has a convergent subsequence which converges to say some \( u^* \in U \). If we take the limit along this subsequence for (3.3), using the assumptions that \( T_n(\cdot|x_n, u_n) \to T(\cdot|x, u) \) weakly, the fact that \( \lim_{n \to \infty} (v_n^*(x_n) - v^*(x)) = c \), and that \( j^*_n \to j^* \) (continuity results from Theorem 3.3.3) we get

\[
\lim_k \left( j^*_{nk} + v^*_{nk}(x_{nk}) \right) = \lim_k \left( c(x, \gamma^*_{nk}(x_{nk})) + \int v^*_{nk}(y) T_{nk}(dy|x_{nk}, \gamma^*_{nk}(x_{nk})) \right) = j^* + v^*(x) + c \approx c(x, u^*) + \int v^*(y) T(dy|x, u^*) + c.
\]

Therefore, \( u^* \) satisfies the ACOE for the kernel \( T \) and thus, any convergent subsequence of \( \gamma^*_n(x_n) \) is an optimal action for \( x \) for the kernel \( T \).

Now consider the following operator \( \hat{T}_n \), for the kernel \( T \) and the policy \( \gamma^*_n \) which is optimal for \( T_n \)

\[
\hat{T}_n \hat{v}_n(x) = c(x, \gamma^*_n(x)) + \int \hat{v}_n(y) T(dy|x, \gamma^*_n(x)).
\]

One can show that this operator is also a contraction under span semi-norm and admits a fixed point \( \hat{v}^*_n \), such that

\[
\hat{j}_n + \hat{v}^*_n(x) = c(x, \gamma^*_n(x)) + \int \hat{v}^*_n(y) T(dy|x, \gamma^*_n(x))
\]

where \( \hat{j}_n = J_\infty(T, \gamma^*_n, x) \) for all \( x \). Hence, we need to show that \( \hat{j}_n \to j^* \) to complete the proof. To show this, in Section 3.8.1 we prove that

\[
\lim_{n \to \infty} \hat{v}^*_n(x_n) - v^*(x) = \hat{c},
\]

(3.7)
3.4. ROBUSTNESS TO INCORRECT CONTROLLED TRANSITION KERNEL MODELS

for any \( x_n \to x \) for some constant \( \hat{c} < \infty \).

Now, assume that \( \lim_n \hat{j}_n \neq j^* \) and that there exists a subsequence \( \hat{j}_{nk} \) and an \( \epsilon > 0 \) such that \( |\hat{j}_{nk} - j^*| > \epsilon \) for every \( k \). We will show that this cannot hold, by establishing the existence of a further subsequence \( \hat{j}_{nk_l} \) which converges to \( j^* \) in the following.

We first note that \( \lim_{n \to \infty} \hat{v}^*_n(x_n) - v^*(x) = \hat{c} \). Hence, [70, Theorem 3.5] (or [97, Theorem 3.5]) yields that 
\[
\int \hat{v}^*_n(y) T(dy|x, \gamma_n) - \hat{v}^*_n(x) \to v^*(y) T(dy|x, u^*) + \hat{c}
\]
also satisfies \( \hat{v}^*_n(x) - v^*(x) \to \hat{c} \).

Therefore, taking the limit along this subsequence,
\[
\lim_{l \to \infty} \hat{j}_{nk_l}
= \lim_{l \to \infty} c(x, \gamma_{nk_l}) + \int \hat{v}^*_{nk_l}(y) T(dy|x, \gamma_n) - \hat{v}^*_n(x)
= c(x, u^*) + \int v^*(y) T(dy|x, u^*) - v^*(x) = j^*.
\]
This contradicts to \( |\hat{j}_{nk} - j^*| > \epsilon \), hence we conclude that \( \hat{j}_n \to j^* \).

3.4.2 Robustness under Setwise Convergence of Transition Kernels

Theorem 3.4.3. We have that \( J_\infty(T, \gamma_n, x) \to J^*(T, x) \) for any \( x \in X \), where \( \gamma_n^* \) is the optimal policy for the transition kernel \( T_n \) that satisfies the ACOE, under Assumption 3.3.2 A, B, C and D if \( T_n(\cdot|x, u_n) \to T(\cdot|x, u) \) weakly for any \( u_n \to u \).

Proof. The proof follows the similar steps as in the proof of Theorem 3.4.2. Consider again the following two ACOE for the kernels \( T_n \) and \( T \) with their fixed point \( v_n^* \) and \( v^* \)
\[
\hat{j}_n^* + v_n^*(x) = \inf_{u \in U} \left[ c(x, u) + \int v_n^*(y) T_n(dy|x, u) \right]
\]
(3.8)
3.4. ROBUSTNESS TO INCORRECT CONTROLLED TRANSITION KERNEL MODELS

\[ j^* + v^*(x) = \inf_{u \in U} \left[ c(x, u) + \int v^*(y) \mathcal{T}(dy|x, u) \right]. \]  

(3.9)

With the same argument used to establish (3.5) now using [64, Theorem 4.8] one show that 
\[ v_n^*(x) - v^*(x) \to c, \]  
for some constant \(|c| < \infty\) for all \(x\).

Since \(U\) is compact, for every \(x\), \(\gamma_n(x)\) has a convergent subsequence which converges to say some \(u^* \in U\). If we take the limit along this subsequence for (3.8), using the assumptions that \(T_n(\cdot|x, u_n) \to T(\cdot|x, u)\) setwise, the fact that \(\lim_{n \to \infty} (v_n^*(x) - v^*(x)) = c\), and that \(j_n^* \to j^*\) (continuity results from Theorem 3.3.3) we get

\[
\lim_k j_{n_k}^* + v_{n_k}^*(x) \\
= c(x, \gamma_{n_k}^*(x)) + \int v_{n_k}^*(y) T_{n_k}(dy|x, \gamma_{n_k}^*(x)) \\
= j^* + v^*(x) + c = c(x, u^*) + \int v^*(y) \mathcal{T}(dy|x, u^*) + c.
\]

Therefore, \(u^*\) satisfies the ACOE for the kernel \(T\) and thus, any convergent subsequence of \(\gamma_n^*(x)\) is an optimal action for \(x\) for the kernel \(T\).

Now consider the operator \(\hat{T}_n\) again,

\[
\hat{T}_n \hat{v}_n(x) = c(x, \gamma_n^*(x)) + \int \hat{v}_n(y) \mathcal{T}(dy|x, \gamma_n^*(x)).
\]

(3.10)

We write

\[
\hat{j}_n + \hat{v}_n^*(x) = c(x, \gamma_n^*(x)) + \int \hat{v}_n^*(y) \mathcal{T}(dy|x, \gamma_n^*(x))
\]

where \(\hat{j}_n = J_\infty(\mathcal{T}, \gamma_n^*, x)\) for all \(x\). Hence, we need to show that \(\hat{j}_n \to j^*\) to complete the proof. By replicating the same arguments in Section 3.8.1 for setwise convergence and
3.4. ROBUSTNESS TO INCORRECT CONTROLLED TRANSITION KERNEL MODELS

using [89, Theorem 20] (setwise convergence with varying functions), one can prove that

\[ \lim_{n \to \infty} \hat{v}_n^*(x) - v^*(x) = \hat{c}, \]

for any \( x \) for some constant \( \hat{c} < \infty \).

Now, assume that \( \lim_n \hat{j}_n \neq j^* \) and that there exists a subsequence \( \hat{j}_{nk} \) and an \( \epsilon > 0 \) such that \( |\hat{j}_{nk} - j^*| > \epsilon \) for every \( k \). We first note that \( \lim_{n \to \infty} \hat{v}_n^*(x) - v^*(x) = \hat{c} \). Hence, [89, Theorem 20] yields that \( \int \hat{v}_{nk}^* (y) T(dy|x, \gamma_{nk}^* (x)) \to \int v^* (y) T(dy|x, u^*) + \hat{c} \) where \( \hat{c} \) also satisfies \( (\hat{v}_{nk}^* (x) - v^*(x)) \to \hat{c} \).

Therefore, taking the limit along this subsequence,

\[
\lim_{l \to \infty} \hat{j}_{nk_l} = \lim_{l \to \infty} c(x, \gamma_{nk_l}^* (x)) + \int \hat{v}_{nk_l}^* (y) T(dy|x, \gamma_{nk_l}^* (x)) - \hat{v}_{nk_l}^* (x) = c(x, u^*) + \int v^* (y) T(dy|x, u^*) - v^*(x) = j^*.
\]

This contradicts to \( |\hat{j}_{nk} - j^*| > \epsilon \), hence we conclude that \( \hat{j}_n \to j^* \). \( \square \)

3.4.3 Robustness under Total Variation Convergence of Transition Kernels

In this section, we will show that for any stationary policy \( \gamma_n^* \) that is optimal for \( \mathcal{T}_n \), as \( \mathcal{T}_n \to \mathcal{T} \) in total variation, we have \( J^*_\infty(\mathcal{T}, \gamma_n^*) \to J^*_\infty(\mathcal{T}) \) under proper conditions.

We note that, since setwise convergence is less stringent than total variation, it is not surprising that we can establish robustness under total variation convergence of the kernels as well. However, in the previous section, we showed convergence only when the policies were restricted to be among those which solves the ACOE for every point in the state space; in the analysis below, the result will be more general and the policies considered are just
3.4. ROBUSTNESS TO INCORRECT CONTROLLED TRANSITION KERNEL MODELS

required to be optimal, without the requirement that they solve the ACOE for every \( x \in X \).

This is not vacuous, as the following example shows.

**Example 3.4.1.** Consider \( X = [-1, 1] \) and \( U = \{0, 1, 2\} \).

Let the kernels be given in the following form for \( n \geq 1 \):

\[
\mathcal{T}_n(\cdot|x,u) = \left( \frac{1}{2} \delta_{\frac{1}{n}}(\cdot) + \frac{1}{2} \delta_{-\frac{1}{n}}(\cdot) \right) 1_{\{x \geq \frac{1}{n}\}} + \left( \frac{1}{2} \delta_{\frac{1}{n}}(\cdot) + \frac{1}{2} \delta_{-\frac{1}{n}}(\cdot) \right) 1_{\{x \leq -\frac{1}{n}\}} + \left( \frac{1}{3} \delta_{\frac{1}{n}}(\cdot) + \frac{1}{3} \delta_{-\frac{1}{n}}(\cdot) + \frac{1}{3} \delta_0(\cdot) \right) 1_{\{-\frac{1}{n} < x < \frac{1}{n}\}}\]

\( \mathcal{T}(\cdot|x,u) = \delta_0(\cdot) \).

The cost function is given by:

\[
c(x,u) = \begin{cases} 
(x)1_{x \geq 0} + 01_{x < 0} & \text{if } u = 0, 1 \\
3 & \text{if } u = 2.
\end{cases}
\]

One can show that this setup satisfies Assumption 3.3.2 A, B, C and D. With this setup, one (among many others) optimal policy for \( \mathcal{T}_n \) when the initial state is \( x = -1 \) is given by:

\[
\gamma_n^*(x) = \begin{cases} 
1 & \text{if } x \leq -\frac{1}{n} \\
0 & \text{if } x \geq \frac{1}{n} \\
2 & \text{otherwise.}
\end{cases}
\]
3.4. ROBUSTNESS TO INCORRECT CONTROLLED TRANSITION KERNEL MODELS

When the initial state is $-1$, the cost under this policy is

$$J_\infty(T_n, \gamma^*_n) = \lim_{N \to \infty} \frac{1}{N} \sum_{t=0}^{N-1} \frac{1}{2} = \frac{1}{2n} \to 0.$$ 

Therefore the policy $\gamma^*_n$ is indeed optimal for $T_n$ for large $n$. An optimal policy for $T$ is given by $\gamma^*(x) = 1$. Thus, the average cost values can be calculated as:

$$J_\infty(T, \gamma^*_n) = \lim_{N \to \infty} \frac{1}{N} \sum_{t=0}^{N-1} E[c(X_t, \gamma^*_n(X_t))]$$

$$= \lim_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N-1} c(0, \gamma^*_n(0)) = \lim_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N-1} 3 = 3$$

Hence, we have that

$$\lim_{n \to \infty} J_\infty(T, \gamma^*_n) = 3 \neq J_\infty(T, \gamma^*) = 0.$$ 

Notice that for this example, $\gamma^*_n$ would not have been be optimal for $T_n$ if the initial state were between $-1/n$ and $1/n$ and, in particular, it does not satisfy the ACOE.

Theorem 3.4.4. We have that $|J_\infty(T, \gamma^*_n) - J_\infty(T)| \to 0$ for any stationary optimal policy $\gamma^*_n$ for $T_n$, under Assumption 3.3.2 A, B, C' and D' if $T_n(\cdot|x, u_n) \to T(\cdot|x, u)$ in total variation for any $u_n \to u$ for every fixed $x$.

Proof. We write:

$$|J_\infty(T, \gamma^*_n) - J_\infty(T)| \leq |J_\infty(T_n) - J_\infty(T)| + |J_\infty(T, \gamma^*_n) - J_\infty(T_n)|$$

the first term goes to by Theorem 3.3.3. For the second term we write

$$|J_\infty(T, \gamma^*_n) - J_\infty(T_n)| \leq \left| J_\infty(T, \gamma^*_n) - \frac{J_t(T_n, \gamma^*_n)}{t} \right|$$
3.5. COMPARISON WITH THE LITERATURE

\[ \frac{J_t(T_n, \gamma_n^*) - J_t(T, \gamma^*)}{t} + \frac{J_t(T, \gamma_n^*) - J_\infty(T, \gamma_n^*)}{t} + \frac{J_t(T, \gamma_n^*) - J_\infty(T, \gamma_n^*)}{t} \]

The first and the last terms above again can be made arbitrarily small by choosing \( t \) large enough uniformly over \( n \) using Lemma 3.3.1. For the second term we use [64, Section A.2] where it is shown that under the stated assumptions \( \sup_{\gamma \in \Gamma} |J_t(T_n, \gamma) - J_t(T, \gamma)| \to 0 \). Hence the proof is complete.

\[ \square \]

3.5 Comparison with the Literature

In the most relevant contribution (to our knowledge), [51], the following problem is considered: Suppose we have an approximating model for the true kernel \( T \), so that, every time step \( t \), an estimate is updated to \( T_t \) and an optimal stationary policy is found with

\[ J_\infty^*(T_t) = \inf_{\gamma \in \Gamma} \lim_{N \to \infty} \frac{1}{N} \sum_{i=0}^{N-1} E^T_t \left[ c(X_i, \gamma(X_i)) \right], \]

where \( E^T_t \) denotes the expectation with transition kernel \( T_t \). Let \( \gamma_t^* \) be the optimal policy for the kernel \( T_t \). It is shown that ([51, Theorem 5.7]) if condition f of Theorem 3.8.1 holds for \( T, T_n \) and if \( \sup_{x,u} \| T_t(\cdot | x, u) - T(\cdot | x, u) \|_{TV} \to 0 \), the following hold:

(i) \( J_\infty(\gamma_t^*) \to J_\infty(\gamma^*) \) as \( t \to \infty \),

(ii) \( J_\infty(T_t, \gamma_t^*) = J_\infty(T, \gamma^*) \).

Notice that the first item (i) is the continuity problem we study in this section (Problem 1). Therefore, it can also be shown to hold true with Theorem 3.3.3 which can be proven with

\[ \sup_u \| T_t(\cdot | x, u) - T(\cdot | x, u) \|_{TV} \to 0 \]
for fixed $x$. Theorem 3.3.3 also states that we can even weaken the total variation convergence of kernels to weak convergence so that it suffices to have $T_t(\cdot|x_t, u_t) \rightarrow T(\cdot|x, u)$ weakly for any $(x_t, u_t) \rightarrow (x, u)$. For the second item (ii), policies $\gamma^*_t$ (since they are not time-invariant) are not stationary for the model $\mathcal{T}$, however it is shown in [51] that using $\gamma^*_t$ at time step $t$, the cost

$$
\limsup_{N \to \infty} \frac{1}{N} \sum_{t=0}^{N-1} E^T[c(X_t, \gamma^*_t(X_t))]
$$

is equal to the optimal cost for the true model.

The following example shows that the total variation convergence of the transition kernels can be too much to ask for deterministic problems. Thus, the relaxation to weak convergence of transition kernels is significant.

**Example 3.5.1.** Assume that $x_0 = 0$ and the transition kernels are given by

$$
\mathcal{T}(\cdot|x) = \delta_0(\cdot), \quad \mathcal{T}_n(\cdot|x) = \delta_{\frac{1}{n}}(\cdot).
$$

The cost function is given as

$$
c(x) = \begin{cases} 
|x| & \text{if } |x| \leq 1, \\
1 & \text{if } |x| > 1.
\end{cases}
$$

Notice that $\|\mathcal{T}_n(\cdot|x) - \mathcal{T}(\cdot|x)\|_{TV} = 2$ for all $n$, however, $\mathcal{T}_n(\cdot|x_n) \rightarrow \mathcal{T}(\cdot|x)$ weakly for any $x_n \rightarrow x$. Furthermore, $J^*_\infty(\mathcal{T}_n) \rightarrow J^*_\infty(\mathcal{T})$. Hence, although the transition kernels do not converge to each other in total variation, continuity still holds. \(\diamondsuit\)
3.6 Robustness under Convergence of Transition Kernels without Uniform Ergodicity

In this section we show that if the family of optimal policies forms an equicontinuous and stationary family we can guarantee continuity and robustness without requiring a uniform ergodicity over policies and initial points. We first present a supporting lemma from [70, Theorem 3.5] and [97, Theorem 3.5].

Lemma 3.6.1. Suppose \( \{ \mu_n \}_n \subset \mathcal{P}(\mathbb{X}) \), where \( \mathbb{X} \) is a Polish space, converges weakly to some \( \mu \in \mathcal{P}(\mathbb{X}) \). For a bounded real valued sequence of functions \( \{ f_n \}_n \) such that \( \| f_n \|_\infty < C \) for all \( n > 0 \) with \( C < \infty \), if \( \lim_{n \to \infty} f_n(x_n) = f(x) \) for all \( x_n \to x \), i.e. \( f_n \) continuously converges to \( f \), then \( \lim_{n \to \infty} \int_{\mathbb{X}} f_n(x) \mu_n(dx) = \int_{\mathbb{X}} f(x) \mu(dx) \).

The following result shows that if we restrict the family of policies to an equicontinuous and stationary family we can guarantee continuity and robustness. For the result we do not require a uniform ergodicity assumption as in Assumption 3.3.1.

Assumption 3.6.1. (i) For any stationary policy \( \gamma \), \( T_n \) and \( T \) lead to positive Harris recurrent chains and in particular admit unique invariant measures \( \pi_{\gamma_n} \) and \( \pi_{\gamma} \).

(ii) \( \{ \pi_{\gamma_n} \}_{\gamma, n} \) and \( \{ \pi_{\gamma} \}_\gamma \) are tight.

(iii) Family of optimal policies \( \{ \gamma_n^* \}_n \) for \( T_n \), is an equicontinuous family of functions and the optimal policy \( \gamma^* \) for \( T \) is continuous.

Theorem 3.6.1. Suppose that Assumption 3.6.1 and Assumption 3.3.2 \( B, C, D \) (for \( T_n \) and \( T \)) hold. Then we have that \( J_{\infty}^*(T_n) \to J^*(T) \) and \( J_{\infty}(T; \gamma_n^*) \to J^*_{\infty}(T) \).

Proof. We now use the following bounds. Let \( \gamma_n^* \) be optimal for \( T_n \) and \( \gamma^* \) be optimal for
3.6. ROBUSTNESS UNDER CONVERGENCE OF TRANSITION KERNELS WITHOUT UNIFORM ERGODICITY

Then,

\[ |J_\infty(T_n, \gamma^*_n) - J_\infty(T, \gamma^*)| \leq \max \left( J_\infty(T_n, \gamma^*) - J_\infty(T_n, \gamma^*_n), J_\infty(T, \gamma^*_n) - J_\infty(T_n, \gamma^*_n) \right), \]

\[ |J_\infty(T, \gamma^*_n) - J_\infty(T)| \leq |J_\infty(T, \gamma^*) - J_\infty(T_n, \gamma^*)| + |J_\infty(T_n, \gamma^*_n) - J_\infty(T, \gamma^*_n)|. \]

(3.11)

Hence, it suffices to show that

\[ |J_\infty(T_n, \gamma^*) - J_\infty(T, \gamma^*)| \to 0, \]

\[ |J_\infty(T, \gamma^*_n) - J_\infty(T_n, \gamma^*_n)| \to 0. \]

First notice that for any policy \( \gamma \), \( J_\infty(T_n, \gamma) = \int c(x) \pi_n^\gamma(dx) \) and \( J_\infty(T, \gamma) = \int c(x) \pi^\gamma(dx) \) for all initial states \( x_0 \in X \) as the chains are positive Harris recurrent. Thus, we only need to show that \( \rho(\pi_n^\gamma, \pi^\gamma) \to 0 \) and \( \rho(\pi_n^\gamma, \pi_n^\gamma) \to 0 \) weakly since \( c \in C_b(X) \), where \( \rho \) metrizes the topology of weak convergence.

For a fixed policy \( \gamma^* \), since \( \gamma^* \) belongs to an equicontinuous family, we have that \( T_n(\cdot|x_n, \gamma^*(x_n)) \to T(\cdot|x, \gamma^*(x)) \) for any \( x_n \to x \) and \( T(\cdot|x, \gamma^*(x)) \) is weakly continuous in \( x \). Since \( \pi_n^\gamma \) is a tight family, there exists a subsequence \( \pi_n^\gamma \) such that \( \pi_n^\gamma \to \pi^\gamma \) weakly for some \( \pi^\gamma \in \mathcal{P}(X) \). As \( \pi_n^\gamma \) is the invariant measure for \( T_n \), we have that for any \( f \in C_b(X) \)

\[ \int f(x_1) T_n(dx_1|x_0, \gamma^*(x_0)) \pi_n^\gamma(dx_0) = \int f(x_0) \pi_n^\gamma(dx_0) \]

using the assumption that \( T_n(\cdot|x_n, \gamma^*(x_n)) \to T(\cdot|x, \gamma^*(x)) \) weakly for any \( x_n \to x \).
and that $\pi_{nk}^{\gamma^*} \to \pi^*$ by taking the limit $k \to \infty$, Lemma 3.6.1 gives us

$$\int f(x_1) T(dx_1|x_0, \gamma^*(x_0))\pi^*(dx_0) = \int f(x_0) \pi^*(dx_0).$$

Since $T$ has a unique invariant measure we can conclude that $\pi^* = \pi^{\gamma^*}$.

So far we have proved that any converging subsequence of $\pi_{n}^{\gamma^*}$ converges weakly to $\pi^{\gamma^*}$. Now suppose that $\pi_{n}^{\gamma^*}$ does not converge to $\pi^{\gamma^*}$. Then, there exists an $\epsilon > 0$ and a further subsequence $\pi_{nk}^{\gamma^*}$ such that $\rho(\pi_{nk}^{\gamma^*}, \pi^{\gamma^*}) > \epsilon$ for all $k$. But because of the tightness assumption there exists a further subsequence $\pi_{nk_l}^{\gamma^*}$ which converges and using the same arguments above it converges to $\pi^{\gamma^*}$ which leads us to a contradiction and completes the proof.

For the sequence of policies $\gamma_{n}^{*}$, using the equicontinuity assumption with Arzelà-Ascoli theorem ([31]), there exists a subsequence $\gamma_{nk}^{*}$ converging uniformly to some $\gamma \in \Gamma$. Therefore, $\mathcal{T}_{nk}(\cdot|x_{nk}, \gamma_{nk}^{*}(x_{nk})) \to \mathcal{T}(\cdot|x, \gamma(x))$ and $\mathcal{T}^{*}(\cdot|x_{nk}, \gamma_{nk}^{*}(x_{nk})) \to \mathcal{T}^{*}(\cdot|x, \gamma(x))$ for any $x_{nk} \to x$. Hence, using the same steps above, we can show that $\pi_{nk}^{\gamma_{nk}^{*}} \to \pi^{\gamma}$ and $\pi^{\gamma_{nk}^{*}} \to \pi^{\gamma}$ weakly. This completes the proof.

**Corollary 3.6.1.** If $\mathcal{X}$ is a finite space, $\mathcal{U}$ is compact, $\mathcal{T}(x, u)$ is weakly continuous in $u$: $c(x, u)$ is continuous in $u$ and under every stationary policy state process $\{x_t\}$ is positive Harris recurrent then we have that $J_{\infty}^{\gamma^{*}}(\mathcal{T}_{n}) \to J^{\gamma}(\mathcal{T})$ and $J_{\infty}(\mathcal{T}, \gamma_{n}^{*}) \to J^{\gamma}_{\infty}(\mathcal{T})$.

**Example 3.6.1 (Adaptive Control).** Suppose a controlled model is given by the dynamics

$$x_{t+1} = Ax_t + Bu_t + w_t$$

where $w_t$ is a i.i.d. Gaussian noise process. Assume that $A$ and $B$ are unknown by the controller. However, the controller can estimate $A$ and $B$ in a consistent way so that
A_n \to A \text{ and } B_n \to B; \text{ with many results reported in the literature [3, 68, 44, 45]. Then, building on Section 2.2.2(i) and 2.2.2(vi), we have } T_n(\cdot|x_n, u_n) \to T(\cdot|x, u) \text{ weakly for any } x_n \to x \text{ and } u_n \to u. \text{ If further we have that the step-wise cost function is in the form } c(x, u) = x^T Q x + u^T R u, \text{ then the optimal policies are linear and also equicontinuous if the model is controllable [9]. Furthermore, since the noise is Gaussian, the chain is Lebesgue irreducible and thus there exists a unique invariant measure and it can be reached from any initial point. Hence, the assumptions of Theorem 3.6.1 are satisfied and the continuity and robustness can be established for this example. In other words; the optimal cost for the estimates } A_n \text{ and } B_n, J^*_\infty(T_n), \text{ converges to optimal cost for } A \text{ and } B, J^*_\infty(T). \text{ Furthermore, if we apply the optimal policies } \gamma_n^* \text{ designed for } A_n \text{ and } B_n, \text{ to the true model } (A, B), \text{ we get } J_\infty(T, \gamma_n^*) \to J^*_\infty(T). \top$

3.7 Application to Empirical Model Learning

3.7.1 Empirical Model Learning and Robustness

Let $T(\cdot|x, u)$ be a transition kernel given previous state and action variables $x \in X, u \in U$, which is unknown to the decision maker (DM). Suppose the DM builds a model for the transition kernels, $T_n(\cdot|x, u)$, for all possible $x \in X, u \in U$ by collecting training data $\{x_t, u_t, \ t \leq n\}$ from the evolving system.

The question we are interested in is that, do we have that the cost calculated under $T_n$ converges to the true cost (i.e., do we have that the cost obtained from applying the optimal policy for the empirical model converges to the true cost as the training length increases)? We will refer to this property as being asymptotically robust under empirical learning. In the following, we provide two setups.
Theorem 3.7.1. Consider the example in Section 2.2.2(vi): Let a controlled model be given as
\[ x_{t+1} = F(x_t, u_t, w_t), \]
where \( \{w_t\} \) is an i.i.d. noise process. Suppose that \( F(x, u, \cdot) : W \to X \) is invertible for all fixed \((x, u)\) and \( F(x, u, w) \) is continuous and bounded on \( X \times U \times W \). If we construct the empirical measures for the noise process \( w_t \) such that for every (fixed) Borel \( B \subset W \), and for every \( n \in \mathbb{N} \), the empirical occupation measures are
\[ \mu_n(B) = \frac{1}{n} \sum_{i=1}^n 1\{F^{-1}_{x_{i-1},u_{i-1}}(x_i) \in B\} \]
where \( F^{-1}_{x_{i-1},u_{i-1}}(x_i) \) denotes the inverse of \( F(x_{i-1}, u_{i-1}, w) : W \to X \) for given \((x_{i-1}, u_{i-1})\). Using the noise measurements, we construct the empirical transition kernel estimates for any \((x_0, u_0)\) and Borel \( B \) as
\[ T_n(B| x_0, u_0) = \mu_n(F^{-1}_{x_0,u_0}(B)). \]

Then, under empirical learning, by Theorem 3.4.2, asymptotic robustness under empirical learning holds.

Proof. By the analysis in Section 2.2.2(vi), we have that
\[ T_n(\cdot|x_n, u_n) \to T(\cdot|x, u), \]
weakly for any \((x_n, u_n) \to (x, u)\) almost surely. Furthermore, additionally let
\[ F(x_t, u_t, w_t) = G(x_t, u_t) + w_t \]
where (in addition to the assumed regularity conditions on \( F \)) we also have that \( G \) has a bounded range and that \( w_t \) has a density which is positive everywhere. Then, by Example 2.2 it follows that Theorem 3.8.1(h) holds. This ensures that under any stationary policy \( x_t \) is positive Harris recurrent and also geometrically ergodic. We thus conclude that under empirical learning, by Theorem 3.4.2 the loss due to an incorrect initial modelling error is zero.

**Theorem 3.7.2.** Suppose we are given the following dynamics for finite state space, \( X \), and finite action space, \( U \),

\[
x_{t+1} = f(x_t, u_t, w_t)
\]

where \( \{w_t\} \) is an i.i.d. noise process and the noise model is unknown. Suppose again that there is an initial training period so that under some policy, every \( x, u \) pair is visited infinitely often if training were to continue indefinitely, but that the training ends at some finite time. Let us assume that, through this training, we empirically learn the transition dynamics such that for every (fixed) Borel \( B \subset X \), for every \( x \in X \), \( u \in U \) and \( n \in \mathbb{N} \), the empirical occupation measures are

\[
\mathcal{T}_n(B|x_0 = x, u_0 = u) = \frac{\sum_{i=1}^{n} 1\{X_i \in B, X_{i-1} = x, U_{i-1} = u\}}{\sum_{i=1}^{n} 1\{X_{i-1} = x, U_{i-1} = u\}}.
\]

Then we have that \( J^*_\beta(\mathcal{T}_n) \rightarrow J^*_\beta(\mathcal{T}) \) and \( J_\beta(\mathcal{T}, \gamma^*_n) \rightarrow J_\beta(\mathcal{T}) \), where \( \gamma^*_n \) is the optimal policy designed for \( \mathcal{T}_n \). Then, asymptotic robustness under empirical learning holds.

**Proof.** By Corollary 2.8.1 \( \mathcal{T}_n(\cdot|x, u) \rightarrow \mathcal{T}(\cdot|x, u) \) weakly for every \( x \in X \), \( u \in U \) almost surely (by law of large numbers). Since the spaces are finite, we also have \( \mathcal{T}_n(\cdot|x, u) \rightarrow \mathcal{T}(\cdot|x, u) \) under total variation.

Suppose further that we have \( \mathcal{T}(\cdot|x, u) > 0 \) for every \( x, u \). Then, for large enough \( n \),
3.7. APPLICATION TO EMPIRICAL MODEL LEARNING

we will have uniform ergodicity by Theorem 3.8.1. As a result, Theorem 3.4.4 will apply and we will have consistency under empirical learning.

3.7.2 Adaptive Learning

Suppose we adaptively learn the transition kernel so that at time \( n \) we have an estimated kernel \( T_n \), and accordingly apply an optimal policy for some time period after \( n \) based on our improved model. In the following, we will build on Theorem 3.4.2, Theorem 3.4.3 or Theorem 3.4.4 to arrive at robustness (and asymptotic consistency) of such an adaptive policy. Let \( T_1, T_2, \cdots \) be a sequence of increasing integers that satisfy

\[
\lim_{k \to \infty} \frac{\sum_{l=1}^{k} T_l}{T_k} = 1.
\]

(3.12)

Suppose that we apply the control policy \( \gamma_{n_k} \), which is optimal for \( T_{n_k} \) with \( n_k = \sum_{l=1}^{k} T_l \), from \( n_k \) until \( n_{k+1} \). Notice that the length of the time intervals we update the policy grows to \( \infty \) with this setup. Call this adaptive policy \( \tilde{\gamma} \). Since the cost is bounded and the rate of convergence to the invariant measure is uniform over policies and over initial states \( x \in X \) under the conditions of ergodicity in Theorem 3.8.1, we have that

\[
J_\infty(T, \tilde{\gamma}, x) = \lim_{T \to \infty} \frac{\sum_{t=0}^{T-1} E_x[c(x_t, \tilde{\gamma}_t(x_t))]}{T} \leq \lim_{k \to \infty} \frac{\sum_{t=n_k}^{n_{k+1}} E_x[c(x_t, \gamma_{n_k}(x_t))]}{n_{k+1}} = \lim_{k \to \infty} J_\infty(T, \gamma_{n_k}, x).
\]

In the above, the last two steps follow from the fact that \( n_{k+1} - n_k \to \infty \) and \( n_{k+1} \to \infty \) at the same rate by (3.12) and that \( \gamma_{n_k} \) is optimal from time \( n_k \) to \( n_{k+1} \). By the results we have established, if \( T_{n_k} \to T \) under any of the senses established by Theorem 3.4.2, Theorem 3.4.3 or Theorem 3.4.4 we will have a strong form of robustness: the cost to due
an incorrect initial modeling error will be zero.

3.8 Technical Results

3.8.1 Proof of Equation (3.7)

We first define the following operator \( \hat{T}_{n,z} \) for some fixed \( z \in X \) by

\[
\hat{T}_{n,z}v(x) := \hat{T}_{n}v(x) - \hat{T}_{n}v(z)
\]

where \( \hat{T}_{n} \) is as in (3.6). We write

\[
\hat{v}^*_n(x_n) - v^*(x) = \left( \hat{v}^*_n(x_n) - \hat{T}^k_{n,z}v^*(x_n) \right) + \left( \hat{T}^k_{n,z}v^*(x_n) - v^*(x) \right)
\]

where \( \hat{T}^k_{n,z} \) is the operator \( \hat{T}_{n,z} \) applied \( k \) consecutive times. We note that \( \hat{T}_{n,z} \) is also a contraction under the span semi-norm. Hence, the first term converges to some \( \hat{c}_1 \) as \( k \to \infty \) uniformly for all \( n \) since \( \hat{T}_{n,z} \) is a contraction uniformly for all \( n \) under the span seminorm and its fixed point is \( \hat{v}^*_n \).

For the second term, we wish to show that \( (\hat{T}^k_{n,z}v^*(x_n) - v^*(x)) \to -v^*(z) \) as \( n \to \infty \) for every fixed \( k < \infty \) for all \( x_n \to x \). We prove this by induction. For \( k = 1 \), we have

\[
\hat{T}_{n,z}v^*(x_n) = c(x_n, \gamma^*_n(x_n)) + \int v^*(y) \mathcal{T}(dy|x_n, \gamma^*_n(x_n)) - \left( c(z, \gamma^*_n(z)) + \int v^*(y) \mathcal{T}(dy|z, \gamma^*_n(z)) \right).
\]

Now assume that \( |\hat{T}^k_{n,z}v^*(x_{n_k}) - v^*(x) + v^*(z)| > \epsilon \) for some \( \epsilon > 0 \) for every \( k \) along
some subsequence $n_k$. We know that there exists a further subsequence, say $n_{k_l}$ along which $\gamma_{n_{k_l}}^*(x_{n_{k_l}}) \to u^*$ and $\gamma_{n_{k_l}}^*(z) \to u^{**}$ for some $u^*, u^{**} \in U$ where $u^*$ is an optimal action for the state $x$ for the kernel $\mathcal{T}$ and $u^{**}$ is optimal for $z$. If we take the limit along this subsequence, i.e.

$$\lim_{l \to \infty} \hat{T}_{n_{k_l}, z} v^*(x_{n_{k_l}})$$

$$= \lim_{l \to \infty} \left( c(x_{n_{k_l}}, \gamma_{n_{k_l}}^*(x_{n_{k_l}})) + \int v^*(y) \mathcal{T}(dy|x_{n_{k_l}}, \gamma_{n_{k_l}}^*(x_{n_{k_l}})) \right)$$

$$- \lim_{l \to \infty} \left( c(z, \gamma_{n_{k_l}}^*(z)) + \int v^*(y) \mathcal{T}(dy|z, \gamma_{n_{k_l}}^*(z)) \right)$$

$$= c(x, u^*) + \int v^*(y) \mathcal{T}(dy|x, u^*)$$

$$- \left( c(z, u^{**}) + \int v^*(y) \mathcal{T}(dy|z, u^{**}) \right)$$

$$= v^*(x) + j^* - v^*(z) - j^* = v^*(x) - v^*(z).$$

Hence we reach a contradiction and it must be that $\hat{T}_{n, z} v^*(x) - v^*(x) \to -v^*(z)$. Now assume that the claim holds for $k$ that is $(\hat{T}_{n, z}^k v^*(x) - v^*(x)) \to -v^*(z)$.

$$\hat{T}_{n, z}^{k+1} v^*(x) = c(x, \gamma_n^*(x)) + \int \hat{T}_{n, z}^k v^*(y) \mathcal{T}(dy|x, \gamma_n^*(x))$$

$$- \left( c(z, \gamma_n^*(z)) + \int \hat{T}_{n, z}^k v^*(y) \mathcal{T}(dy|z, \gamma_n^*(z)) \right).$$

If we take a subsequence, indexed by say $m$, along which $\gamma_m^*(x_m) \to u^*$ and $\gamma_m^*(z) \to u^{**}$ where $u^*$ is an optimal action for $x$ and $u^{**}$ is optimal for $z$, then by taking the limit along this subsequence

$$\lim_{m \to \infty} \hat{T}_{m}^{k+1} v^*(x_m)$$
3.8. TECHNICAL RESULTS

\[ = \lim_{m \to \infty} c(x_m, \gamma_m^*(x_m)) + \int \hat{T}_{m}^k \nu^*(y) \mathcal{T}(dy|x_m, \gamma_m^*(x_m)) \]
\[ - \lim_{m \to \infty} c(z, \gamma_m^*(z)) + \int \hat{T}_{m}^k \nu^*(y) \mathcal{T}(dy|z_m, \gamma_m^*(z)) \]
\[ = c(x, u^*) + \int (\nu^*(y) - \nu^*(z)) \mathcal{T}(dy|x, u^*) \]
\[ - \left( c(z, u^{**}) + \int (\nu^*(y) - \nu^*(z)) \mathcal{T}(dy|z, u^{**}) \right) \]
\[ = \nu^*(x) - \nu^*(z). \]

Hence, a similar contradiction argument, we used for the case \( k = 1 \) yields that \( (\hat{T}_{k}^n \nu^*(x_n) - \nu^*(x)) \to -\nu^*(z) \) as \( n \to \infty \) for every fixed \( k < \infty \) for all \( x_n \to x \). Thus, we have that

\[ \hat{\nu}_n^*(x_n) - \nu^*(x) = (\hat{\nu}_n^*(x_n) - \hat{T}_{n,z}^k \nu^*(x_n)) \]
\[ + (\hat{T}_{n,z}^k \nu^*(x_n) - \nu^*(x)) \to \hat{c}_1 - \nu^*(z) := \hat{c}. \]

3.8.2 Ergodicity conditions on Markov Chains

The following theorem is stated here for easy reference.

**Theorem 3.8.1.** [54 Theorem 3.2] Consider the following.

a. There exists a state \( x^* \in X \) and a number \( \beta > 0 \) such that \( \mathcal{T}(\{x^*\} | x, \gamma) \geq \beta \), for all \( x \in X, \gamma \in \Gamma_s \).

b. There exists a positive integer \( t \) and a non-trivial measure \( \mu \) on \( X \) such that \( \mathcal{T}^t(\cdot | x, \gamma) \geq \mu(\cdot) \) for all \( x \in X, \gamma \in \Gamma_s \).

c. For each \( \gamma \in \Gamma_s \), the transition kernel \( \mathcal{T}(dy|x, \gamma) \) has a density \( p(y|x, \gamma) \) with respect to a sigma-finite measure \( m \) on \( X \), and there exist \( \epsilon > 0 \) and \( C \in \mathcal{B}(X) \) such that \( m(C) > 0 \) and \( p(y|x, \gamma) \geq \epsilon \) for all \( y \in C, x \in X, \gamma \in \Gamma_s \).
d. For each $\gamma \in \Gamma_s$, $T(dy|x,\gamma)$ has a density $p(y|x,\gamma)$ with respect to a sigma-finite measure $m$ on $X$, and $p(y|x,\gamma) \geq p_0(y)$ for all $x, y \in X$, $\gamma \in \Gamma_s$, where $p_0$ is a non-negative measurable function with $\int p_0(y)m(dy) > 0$.

e. There exists a positive integer $t$ and a measure $\mu$ on $X$ such that $\mu(X) < 2$ and $T^t(\cdot|x,\gamma) \leq \mu(\cdot)$ for all $x \in X$, $\gamma \in \Gamma_s$.

f. There exists a positive integer $t$ and a positive number $\beta < 1$ such that $\|T^t(\cdot|x,\gamma) - T^t(\cdot|x',\gamma)\|_{TV} \leq 2\beta$ for all $x, x' \in X$, $\gamma \in \Gamma_s$.

g. There exists a positive integer $t$ and a positive number $\beta$ for which the following holds: For each $\gamma \in \Gamma_s$, there is a probability measure $\mu_\gamma$ on $X$ such that $T^t(\cdot|x,\gamma) \geq \beta \mu_\gamma(\cdot)$ for all $x \in X$.

h. There exist positive numbers $c$ and $\beta$, with $\beta < 1$, for which the following holds: For each $\gamma \in \Gamma_s$, there is a probability measure $p_\gamma$ on $X$ such that $\|T^t(\cdot|x,\gamma) - p_\gamma(\cdot)\|_{TV} \leq c\beta^t$ for all $x \in X$, $t \in \mathbb{N}$.

i. The state process is uniformly ergodic such that $\lim_{t \to \infty} \|T^t(\cdot|x,\gamma) - p_\gamma(\cdot)\|_{TV} = 0$ uniformly in $x \in X$ and $\gamma \in \Gamma_s$.

The conditions above are related as follows:

\[\begin{align*}
a &\rightarrow b \\
e &\rightarrow f \\
c &\rightarrow d \rightarrow b \rightarrow f \leftrightarrow g \leftrightarrow h \leftrightarrow i.\end{align*}\]
Chapter 4

Regularity Relations Between POMDPs and Their Fully Observed Belief MDP Reduction

Weak continuity of the transition kernel of a controlled or a control-free Markov chain is a commonly imposed assumption for the analysis of Markov chains. In particular, for our robustness results in Chapter 2 and 3, we made use of the weak continuity of transition kernels for several key results. Hence, in this chapter, we present sufficient conditions on a partially observed model that make the transition kernel of the corresponding reduced fully observed model, weakly continuous (known as the non-linear filter kernel as a probability-measure valued Markovian system).

4.1 Markov Property of Filter Processes

We start with reviewing the probabilistic setup of the partially observed Markov processes. For this section, we will consider a more general channel version where the channel also depends on the control action variable. As before, we let $\mathbb{X} \subset \mathbb{R}^n$ be a Borel set in which a control-free or controlled Markov process $\{X_t, t \in \mathbb{Z}_+\}$ takes its value. Let $U$, the action space, be a Borel subset of some Euclidean space $\mathbb{R}^p$. Let $\mathbb{Y} \subset \mathbb{R}^m$ be a Borel set, and let
an observation channel \( Q \) be defined as a stochastic kernel (regular conditional probability) from \( X \times U \) to \( Y \) such that \( Q(\cdot|x,u) \) is a probability measure on the Borel \( \sigma \)-algebra \( B(Y) \) of \( Y \) for every \( (x,u) \in X \times U \) and \( Q(A|\cdot) : X \times U \to [0,1] \) is a Borel measurable function for every \( A \in B(Y) \). Note that, in general, we have and will assume that the channel \( Q \) only depends on the most recent state variable and not on the control action variable, however, only for this chapter, we will consider a more general channel model which also depends on the control.

Let a decision maker (DM) be located at the output of an observation channel \( Q \), with inputs \( (X_t, U_{t-1}) \) and outputs \( Y_t \). An admissible policy \( \gamma \) is a sequence of control functions \( \{\gamma_t, t \in \mathbb{Z}_+\} \) such that \( \gamma_t \) is measurable with respect to the \( \sigma \)-algebra generated by the information variables

\[
I_t = \{Y_{[0,t]}, U_{[0,t-1]}\}, \quad t \in \mathbb{N}, \quad I_0 = \{Y_0\},
\]

where

\[
U_t = \gamma_t(I_t), \quad t \in \mathbb{Z}_+ \tag{4.1}
\]

are the \( U \)-valued control actions and

\[
Y_{[0,t]} = \{Y_s, 0 \leq s \leq t\}, \quad U_{[0,t-1]} = \{U_s, 0 \leq s \leq t - 1\}.
\]

We define \( \Gamma \) to be the set of all such admissible policies.

The joint distribution of the state, control, and observation processes is determined the following system dynamics:

\[
\Pr((X_0, Y_0) \in B) = \int_B Q_0(dy_0|x_0)P_0(dx_0), B \in B(X \times Y),
\]
where $P_0$ is the prior distribution of the initial state $X_0$ and $Q_0$ is the prior control-free observation channel, and for $t \in \mathbb{N}$

$$
\text{Pr}
\begin{pmatrix}
(X_t, Y_t) \\
\end{pmatrix}
\in B
\left|
\begin{pmatrix}
(X, Y, U)_{[0,t-1]} = (x, y, u)_{[0,t-1]}
\end{pmatrix}
\right.
= 
\int_B Q(dy|_{x_t, u_{t-1}}) \mathcal{T}(dx|_{x_{t-1}, u_{t-1}}),
$$

where $\mathcal{T}(\cdot|x, u)$ is a stochastic kernel from $X \times U$ to $X$. This completes the probabilistic setup of the partially observed models.

It is known that any partially observed Markov process (POMP) can be reduced to a completely observable Markov process ([116], [88]), whose states are the posterior state distributions or 'beliefs' of the observer; that is, the state at time $t$ is

$$
Z_t(\cdot) := \text{Pr}\{X_t \in \cdot | Y_0, \ldots, Y_t, U_0, \ldots, U_{t-1}\} \in \mathcal{P}(X).
$$

We call this equivalent process the filter process. The filter process has state space $Z = \mathcal{P}(X)$ and action space $U$. Note that $Z$ is equipped with the Borel $\sigma$-algebra generated by the topology of weak convergence [14]. Under this topology, $Z$ is a standard Borel space [82]. The transition probability of the filter process can be constructed as follows.

The joint conditional probability on next state and observation variables given the current control action and the current state of the filter process is given by

$$
R(B \times C|u_0, z_0) = \int_X \int_B Q(C|x_1, u_0) \mathcal{T}(dx_1|x_0, u_0)z_0(dx_0),
$$

for all $B \in \mathcal{B}(X)$ and $C \in \mathcal{B}(Y)$. Then, the conditional distribution of the next observation variable given the current state of the filter process and the current control action is given
4.1. MARKOV PROPERTY OF FILTER PROCESSES

by

\[ P(C|u_0, z_0) = \int_X \int_X Q(C|x_1, u_0) T(dx_1|x_0, u_0) z_0(dx_0), \]

for all \( C \in B(Y) \). Using this, we can disintegrate \( R \) (see [10, Proposition 7.27]) as follows:

\[ R(B \times C|u_0, z_0) = \int_C F(B|y_1, u_0, z_0) P(dy_1|u_0, z_0) \]
\[ = \int_C z_1(y_1, u_0, z_0)(B) P(dy_1|u_0, z_0), \quad (4.3) \]

where \( F \) is a stochastic kernel from \( Z \times Y \times U \) to \( X \) and the posterior distribution of \( x_1 \), determined by the kernel \( F \), is the state variable \( z_1 \) of the filter process. Then, the transition probability \( \eta \) of the filter process can be constructed as follows. If we define the measurable function \( F(z,u,y) := F(\cdot|y,u,z) = \Pr \{ X_{t+1} \in \cdot | Z_t = z, U_t = u, Y_{t+1} = y \} \) from \( Z \times U \times Y \) to \( Z \) and use the stochastic kernel \( P(\cdot|z,u) = \Pr \{ Y_{t+1} \in \cdot | Z_t = z, U_t = u \} \) from \( Z \times U \) to \( Y \), we can write \( \eta \) as

\[ \eta(\cdot|z,u) = \int_Y 1_{\{F(z,u,y)\in \cdot\}} P(dy|z,u). \quad (4.4) \]

The one-stage cost function \( \tilde{c} : Z \times U \rightarrow [0, \infty) \) of the filter process is given by

\[ \tilde{c}(z,u) := \int_X c(x,u) z(dx), \quad (4.5) \]

which is a Borel measurable function. Hence, the filter process is a completely observable Markov process with the components \( (Z, U, \tilde{c}, \eta) \).
4.2 PROBLEM FORMULATION

For the filter process, the information variables is defined as

\[ \tilde{I}_t = \{ Z_{[0,t]}, U_{[0,t-1]} \}, \quad t \in \mathbb{N}, \quad \tilde{I}_0 = \{ Z_0 \}. \]

It is a standard result that an optimal control policy of the original POMP will use the belief $Z_t$ as a sufficient statistic for optimal policies (see [116], [88]). More precisely, the filter process is equivalent to the original POMP in the sense that for any optimal policy for the filter process, one can construct a policy for the original POMP which is optimal.

4.2 Problem Formulation

We will first review some metrics for spaces of probability measures. As we have discussed in Section 2.2, the topology of weak convergence on the set of probability measures on a separable metric space $S$ is metrizable. One such metric is the bounded-Lipschitz metric. For any two probability measures $\mu$ and $\nu$, the bounded-Lipschitz metric is defined as:

\[
\rho_{BL}(\mu, \nu) = \sup_{\|f\|_{BL} \leq 1} \left\| \int_S f(x) \mu(dx) - \int_S f(x) \nu(dx) \right\| \tag{4.6}
\]

where $\|f\|_{BL} = \|f\|_{\infty} + \sup_{x \neq y} \frac{|f(x) - f(y)|}{d(x,y)}$ and $\|f\|_{\infty} = \sup_{x \in S} |f(x)|$. Another metric that metrizes the weak topology on $\mathcal{P}(S)$ is the following:

\[
\rho(\mu, \nu) = \sum_{m=1}^{\infty} 2^{-(m+1)} \left| \int_S f_m(x) \mu(dx) - \int_S f_m(x) \nu(dx) \right|
\]

where $\{f_m\}_{m \geq 1}$ is an appropriate sequence of continuous and bounded functions such that $\|f_m\|_{\infty} \leq 1$ for all $m \geq 1$ (see [82] Theorem 6.6, p. 47).

Another metric, we will use in this chapter is the total variation distance (see Section
Recall that for probability measures \( \mu, \nu \in \mathcal{P}(S) \), the total variation distance between \( \mu \) and \( \nu \) is given by

\[
\|\mu - \nu\|_{TV} = 2 \sup_{B \in \mathcal{B}(S)} |\mu(B) - \nu(B)| = \sup_{f: \|f\|_\infty \leq 1} \left| \int f(x) \mu(dx) - \int f(x) \nu(dx) \right|,
\]

where the supremum is taken over all Borel measurable real \( f \) such that \( \|f\|_\infty \leq 1 \).

**Definition 4.2.1 (Weak Feller Property).** We say that a Markov decision process with transition kernel \( T(\cdot|x, u) \) has weak Feller property if \( T \) is weakly continuous in \( x \) and \( u \); that is, if \( (x_n, u_n) \to (x, u) \), then \( T(\cdot|x_n, u_n) \to T(\cdot|x, u) \) weakly.

With this definition, we can now state the problem that we are interested in this chapter.

**(P)** Under what conditions on the transition kernel and the observation channel of the partially observed model, the filter process has a weak Feller property?

### 4.3 Significance of the Problem

In this section, we motivate the operational (in view of engineering applications) and the mathematical importance of the problem in view of existence and invariance properties, approximations and computational results involving non-linear filters and stochastic control, and related applications involving particle filtering.

For finite-horizon problems and a large class of infinite-horizon discounted cost problems, it is a standard result that an optimal control policy will use the filter process as a sufficient statistic for optimal policies (see [116, 88, 15]). Hence, the partially observed model and the corresponding filter process are equivalent in the sense of cost minimization. Therefore, results developed for the standard controlled Markov process problems (e.g., measurable selection criteria as summarized in [52, Chapter 3]) can be applied to the
4.3. SIGNIFICANCE OF THE PROBLEM

filter processes, and so, to the partially observed models. In controlled Markov processes theory, weak continuity of the transition kernel is an important condition leading to both the existence of optimal control policies for finite-horizon and infinite-horizon discounted cost problems as well as the continuity properties of the value function (see, e.g., [53, Section 8.5]).

For partially observed stochastic control problems with the average cost criterion, the conditions of existence of optimal policies stated in the literature are somewhat restrictive, with the most relaxed conditions to date being reported in [19, 22], to our knowledge. For such average cost stochastic control problems, the weak Feller property can be utilized to obtain a direct method to establish the existence of optimal stationary (and possibly randomized) control policies. Indeed, for such problems, the convex analytic method ([74] and [18]) is a powerful approach to establish the existence of optimal policies. If one can establish the weak Feller property of the filter process, then the continuity and compactness conditions utilized in the convex program of [18] would lead to the existence of optimal average cost policies.

In addition to existence of optimal policies, the weak Feller property has also recently been shown to lead to (asymptotic) consistency in approximation methods for controlled Markov processes with uncountable state and action spaces. In [92, 93], authors showed that optimal policies obtained from finite-model approximations to infinite-horizon discounted cost controlled Markov processes with Borel state and action spaces asymptotically achieve the optimal cost for the original problem under weak Feller property. Hence, the weak Feller property of filter process suggests that approximation results in [92, 93], which only require weak continuity conditions on the transition probability of a given controlled Markov model, are particularly suitable in developing approximation methods for
partially observed models (through their reduction to fully observed models).

For control-free setups, the weak Feller property of the filter processes leads to the existence of invariant probability measures when the hidden Markov processes take their values in compact spaces or more general spaces under appropriate tightness/regularity conditions [71, 23].

For the empirical consistency and convergence results involving the very popular particle filtering algorithms, weak Feller property of the filter processes is a commonly imposed, implicit, assumption (see e.g. [27, 26]). Finally, for the study of ergodicity and asymptotic stability of nonlinear filters, weak Feller property also plays an important role (see [100, 12, 23]).

4.4 Main Results and Connections with the Literature

4.4.1 Statement of Main Results

In this section, we show the weak Feller property of the filter process under two different set of assumptions.

Assumption 4.4.1.

(i) The transition probability $\mathcal{T}(\cdot | x, u)$ is weakly continuous in $(x, u)$, i.e., for any $(x_n, u_n) \to (x, u)$, $\mathcal{T}(\cdot | x_n, u_n) \to \mathcal{T}(\cdot | x, u)$ weakly.

(ii) The observation channel $Q(\cdot | x, u)$ is continuous in total variation, i.e., for any $(x_n, u_n) \to (x, u)$, $Q(\cdot | x_n, u_n) \to Q(\cdot | x, u)$ in total variation.

Assumption 4.4.2.
4.4. MAIN RESULTS AND CONNECTIONS WITH THE LITERATURE

(i) The transition probability $T(\cdot|x,u)$ is continuous in total variation in $(x,u)$, i.e., for any $(x_n, u_n) \to (x,u)$, $T(\cdot|x_n, u_n) \to T(\cdot|x,u)$ in total variation.

(ii) The observation channel $Q(\cdot|x)$ is independent of the control variable.

We now formally state the main results of this section.

**Theorem 4.4.1** (Feinberg et. al. [42]). Under Assumption 4.4.1, the transition probability $\eta(\cdot|z,u)$ of the filter process is weakly continuous in $(z,u)$.

**Theorem 4.4.2.** Under Assumption 4.4.2, the transition probability $\eta(\cdot|z,u)$ of the filter process is weakly continuous in $(z,u)$.

Theorem 4.4.1 is originally due to Feinberg et. al. [42]. The contribution of us is that the proof presented here is more direct and significantly more concise. Theorem 4.4.2 establishes that under Assumption 4.4.2 (with no assumptions on the measurement model), the filter process is weakly continuous. This result has not been previously reported in the literature.

In addition these weak continuity results, we also present the following result which establishes the Lipschitz continuity of $\eta$ under the $\rho_{BL}$ metric. This result will be useful for Chapter 5.

**Theorem 4.4.3.** Assume that

$$\| T(\cdot|x,u) - T(\cdot|x',u) \|_{TV} \leq \alpha_X |x - x'|$$

for some $\alpha_X < \infty$ for all $u \in U$. Assume also that the observation channel $Q(\cdot|x)$ is
independent of the control variable. We have

\[ \rho_{BL}(\eta(\cdot|z,u),\eta(\cdot|z',u)) \leq 5(1 + \alpha \chi)\rho_{BL}(z,z') \]

where \( \rho_{BL} \) is the bounded Lipschitz metric that metrizes the weak convergence topology, which is defined for \( \mu, \nu \in \mathcal{P}(X) \) as

\[ \rho_{BL}(\mu, \nu) := \sup_{\|f\|_{BL} \leq 1} \left| \int f \, d\mu - \int f \, d\nu \right| \]

where \( \|f\|_{BL} = \|f\|_{\infty} + \sup_{x \neq y} \frac{|f(x) - f(y)|}{d(x,y)} \) and \( \|f\|_{\infty} = \sup_{x \in X} |f(x)| \).

The proofs of these results are presented in Section 4.4.4.

4.4.2 Examples

In this section, we give concrete examples for the system and observation channel models which satisfy Assumption 4.4.1 or Assumption 4.4.2. Suppose that the system dynamics and the observation channel are represented as follows:

\[ X_{t+1} = H(X_t, U_t, W_t), \]
\[ Y_t = G(X_t, U_{t-1}, V_t), \]

where \( W_t \) and \( V_t \) are i.i.d. noise processes. This is, without loss of generality, always the case; that is, one can transform the dynamics of any partially observed model into this form (see Lemma 1.2 in [43], or Lemma 3.1 of [17]).

(i) Suppose that \( H(x, u, w) \) is a continuous function in \( x \) and \( u \). Then, the corresponding transition kernel is weakly continuous. To see this, observe that, for any \( c \in C_b(X) \),
we have

\[
\int \mathcal{C}(x_1) T(dx_1|x_0^n, u_0^n) = \int c(H(x_0^n, u_0^n, w_0)) \mu(dw_0)
\]

\[
\to \int c(H(x_0, u_0, w_0)) \mu(dw_0) = \int c(x_1) T(dx_1|x_0, u_0),
\]

where we use \( \mu \) to denote the probability model of the noise.

(ii) Suppose that \( G(x, u, v) = g(x, u) + v \), where \( g \) is a continuous function and \( V_t \) admits a continuous density function \( \varphi \) with respect to some reference measure \( \nu \). Then, the channel is continuous in total variation. Notice that under this setup, we can write

\[ Q(dy|x, u) = \varphi(y - h(x, u)) \nu(dy). \]

Thus, the density of \( Q(dy|x_n, u_n) \) converges to the density of \( Q(dy|x, u) \) pointwise, and so, \( Q(dy|x_n, u_n) \) converges to \( Q(dy|x, u) \) in total variation by Scheffé’s Lemma [13, Theorem 16.12]. Hence, \( Q(dy|x, u) \) is continuous in total variation under these conditions.

(iii) Suppose that we have \( H(x, u, w) = h(x, u) + w \), where \( f \) is continuous and \( W_t \) admits a continuous density function \( \varphi \) with respect to some reference measure \( \nu \). Then, the transition probability is continuous in total variation. Again, notice that with this setup we have

\[ T(dx_1|x_0, u_0) = \varphi(x_1 - h(x_0, u_0)) \nu(dx_1). \]

Thus, continuity of \( \varphi \) and \( h \) guarantees the pointwise convergence of the densities, so we can conclude that the transition probability is continuous in total variation by again Scheffé’s Lemma.

The analysis in this section will provide weak Feller results for a large class of partially observed control systems as reviewed in the aforementioned examples. In particular, if the state dynamics are affected by an additive noise process which admits a continuous density, we can guarantee weak Feller property of the filter process by means of Theorem 4.4.2 without referring to the noise model of the observation channel.
4.4. MAIN RESULTS AND CONNECTIONS WITH THE LITERATURE

4.4.3 Comparison with the Prior Results

Weak Feller property of the control-free transition probability of the filter processes has been established using different approaches and different conditions. In [12] it has been shown that, for continuous-time filter processes, if the signal process (state process of the partially observed model) is weak Feller and the measurement channel is an additive channel in the form \( Y_t = \int_0^t h(X_u)du + V_t \), where \( h \) is assumed to be continuous and possibly unbounded and \( V_t \) is a standard Wiener process, then the filter process itself is also weak Feller. In [23], the authors study the discrete-time filter processes, where the state process noise may not be independent of the observation process noise; it has been shown that if the observation model is additive in the form \( Y_t = h(X_t) + V_t \), where \( h \) is assumed to be continuous and \( V_t \) is an i.i.d. noise process which admits a continuous and bounded density function, then the observation and filter processes \((Y_t, Z_t)\) are jointly weak Feller. In [100], the weak Feller property of the filter process has been shown for both discrete and continuous time setups when the channel is additive, \( Y_t = h(X_t) + V_t \), where \( h \) is bounded and continuous and \( V_t \) is an i.i.d. noise process with a continuous, bounded and positive density function.

[27, 26] have studied the consistency of the particle filter methods where the weak Feller property of the filter process has been used to establish the convergence results. In [27], it has been shown that the filter process is weak Feller under the assumption that the transition probability of the partially observed system is weak Feller and the measurement channel is an additive channel in the form \( Y_t = h(X_t) + V_t \), where \( h \) is assumed to be continuous and \( V_t \) is an i.i.d. noise process, which admits a continuous and bounded density function with respect to the Lebesgue measure. In [26], the weak Feller property of the filter process has been established under the assumption that the measurement channel admits a continuous
and bounded density function with respect to the Lebesgue measure; i.e., the channel can be written in the following form: 
\[ Q(y \in A|x) = \int_A g(x, y)dy \]
for any \( A \in \mathcal{B}(\mathbb{Y}) \) and for any \( x \in X \), where \( g \) is a continuous and bounded function.

Weak Feller property of the controlled transition probability of the filter processes has been established, in the most general case to date, by Feinberg et.al. [42]. Under the assumption that the measurement channel is continuous in total variation and the transition kernel of the partially observed model is weak Feller, the authors have established the weak Feller property of the transition probability of the filter process. In Section 4.4.5, we will give a detailed discussion on the methods used by Feinberg et.al. [42], and also, we will compare their approach with ours.

As reviewed above, the prior literature often used the additive channel model 
\[ Y_t = h(X_t, U_{t-1}) + V_t \]
with various regularity assumptions on \( h \) and the noise model \( V_t \). When the observation channel is additive \( Y_t = h(X_t, U_{t-1}) + V_t \), where \( h \) is continuous and \( V_t \) admits a continuous density function with respect to some measure \( \mu \), one can show that the channel also admits a continuous density function, i.e., 
\[ Q(y \in A|x, u) = \int_A g(x, u, y)\mu(dy) \]
for any \( A \in \mathcal{B}(\mathbb{Y}) \) and for any \((x, u)\) \( \in X \times U \). When the observation channel has a continuous density function, the pointwise convergence of the density functions implies the total variation convergence by Scheffé’s Lemma [13, Theorem 16.12]. Thus, 
\[ g(x_k, u_k, y) \rightarrow g(x, u, k) \]
for some \( (x_k, u_k) \rightarrow (x, u) \) implies that \( Q(\cdot|x_k, u_k) \rightarrow Q(\cdot|x, u) \) in total variation, i.e., the observation channel is continuous in total variation.

In the following, we develop a relationship between the total variation continuity of the channel (as required by [42] and in our Theorem 4.4.1) and the more restrictive density conditions on the measurement channels presented in the prior works [12, 23, 26, 27].

In the theorem below, we show that having a continuous density is almost equivalent to
the condition that the observation channel is continuous in total variation. To our knowledge, it is the first result in the literature making the connection between channels which are continuous in total variation and channels which admit a density function.

**Theorem 4.4.4.** Suppose that the observation channel $Q(dy|x,u)$ is continuous in total variation. Then, for any $(z,u) \in Z \times U$, we have, $\mathcal{T}(\cdot|z,u)$-a.s., that $Q(dy|x,u) \ll P(dy|u,z)$ and

$$Q(dy|x,u) = g(x,u,y)P(dy|z,u)$$

for a measurable function $g$, which satisfies for any $A \in \mathcal{B}(Y)$ and for any $x_k \to x$

$$\int_A |g(x_k,u,y) - g(x,u,y)|P(dy|z,u) \to 0.$$ 

**Proof.** Fix any $(z,u)$. We first show that $Q(dy|x,u) \ll P(dy|z,u)$, $\mathcal{T}(\cdot|z,u)$-a.s.. Note that $Q(dy|x,u) \ll P(dy|z,u)$ if and only if, for all $\varepsilon > 0$, there exists $\delta > 0$ such that $Q(A|x,u) < \varepsilon$ whenever $P(A|z,u) < \delta$. For each $n \geq 1$, let $K_n \subset X$ be compact such that $\mathcal{T}(K_n|z,u) > 1 - 1/3n$. As $Q(dy|x,u)$ is continuous in total variation norm, the image of $K_n \times \{u\}$ under $Q(dy|x,u)$ is compact in $\mathcal{P}(Y)$. Hence, there exist $\{\nu_1, \ldots, \nu_l\} \subset \mathcal{P}(Y)$ such that

$$\max_{x \in K_n} \min_{i=1,\ldots,l} \|Q(\cdot|x,u) - \nu_i\|_{TV} < 1/3n.$$

Define the following stochastic kernel $\nu_n(\cdot|x,u) = \arg \min_{\nu_i} \|Q(\cdot|x,u) - \nu_i\|_{TV}$. Then, we define $P_n(\cdot|z,u) = \int_X \nu_n(\cdot|x,u)\mathcal{T}(dx|z,u)$. One can prove that $\|P(\cdot|z,u) - P_n(\cdot|z,u)\|_{TV} < 1/n$. Moreover, since $P_n(\cdot|z,u)$ is a mixture of finite probability measures $\{\nu_1, \ldots, \nu_l\}$, we
have that $\nu_n(\cdot|x, u) \ll P_n(\cdot|z, u)$ for all $x \in C_n$, where $T(C_n|z, u) = 1$. Let $C = \bigcap_n C_n$, and so, $T(C|z, u) = 1$. We claim that if $x \in C$, then $Q(dy|x, u) \ll P(dy|z, u)$, which completes the proof of the first statement. To prove the claim, fix any $\varepsilon > 0$ and choose $n \geq 1$ such that $\varepsilon > 2/3n$. Then, there exists $\delta > 0$ such that $\nu_n(A|x, u) < \varepsilon/2$ whenever $P_n(A|z, u) < \delta$. This implies that $Q(A|x, u) < \varepsilon$ whenever $P(A|z, u) < \delta + 1/n$. Hence, $Q(dy|x, u) \ll P(dy|z, u)$.

To show the second claim, for any $A \in B(\mathcal{Y})$ and for any $x_k \to x$, we define

$$A_{+}^{(k)} := \{y \in A : g(x_k, u, y) > g(x, u, y)\},$$

$$A_{-}^{(k)} := \{y \in A : g(x_k, u, y) < g(x, u, y)\}.$$

With these sets, we have

$$\int_A |g(x_k, u, y) - g(x, u, y)| P(dy|z, u)$$

$$= \int_{A_{+}^{(k)}} g(x_k, u, y) P(dy|z, u) - \int_{A_{-}^{(k)}} g(x_k, u, y) P(dy|z, u)$$

$$+ \int_{A_{+}^{(k)}} g(x, u, y) P(dy|z, u) - \int_{A_{-}^{(k)}} g(x, u, y) P(dy|z, u)$$

$$\leq |Q(A_{+}^{(k)}|x_k, u) - Q(A_{+}^{(k)}|x, u)| + |Q(A_{-}^{(k)}|x_k, u) - Q(A_{-}^{(k)}|x, u)|$$

$$\leq 2\|Q(\cdot|x_k, u) - Q(\cdot|x, u)\|_{TV} \to 0.$$

We again emphasize that weak Feller property of the filter process under Assumption 4.4.1 has been first established by [42] using different method compared to ours.
method is significantly more concise and direct. It is also important to note that Assumption 4.4.2 completely eliminates any restriction on the observation channel to establish the weak Feller property of filter process. This relaxation is quite important in practice since modelling the noise on the observation channel in control problems is quite cumbersome, and in general, infeasible. But, in many problems that arise in practice, the transition probability has a continuous density with respect to some reference measure, which directly implies, via Scheffé’s Lemma, the total variation continuity of the transition probability. We also note that the weak Feller property under only Assumption 4.4.2-(i) cannot be established. Indeed, Example 4.1 of [42] shows that the total variation continuity assumption on the observation channel cannot be relaxed even when the transition kernel is continuous in total variation to prove weak Feller property of the filter process under a controlled observation channel model: A careful look at the counterexample shows that it indeed uses the discontinuity of the observation channel in the control action to prove that the filter process cannot be weak Feller when the observation channel is not continuous in total variation and the transition kernel is continuous in total variation.

4.4.4 Proofs

The following result will play a key role for the proofs of main results.

**Lemma 4.4.1.** Let $X$ be a Borel space. Suppose that we have a family of uniformly bounded real Borel measurable functions $\{f_{n,\lambda}\}_{n \geq 1, \lambda \in \Lambda}$ and $\{f_{\lambda}\}_{\lambda \in \Lambda}$, for some set $\Lambda$. If, for any $x_n \to x$ in $X$, we have

$$\lim_{n \to \infty} \sup_{\lambda \in \Lambda} |f_{n,\lambda}(x_n) - f_{\lambda}(x)| = 0 \quad (4.7)$$
4.4. MAIN RESULTS AND CONNECTIONS WITH THE LITERATURE

\[
\lim_{n \to \infty} \sup_{\lambda \in \Lambda} |f_\lambda(x_n) - f_\lambda(x)| = 0, \quad (4.8)
\]

then, for any \( \mu_n \to \mu \) weakly in \( \mathcal{P}(X) \), we have

\[
\lim_{n \to \infty} \sup_{\lambda \in \Lambda} \left| \int_X f_{n,\lambda}(x) \mu_n(dx) - \int_X f_\lambda(x) \mu(dx) \right| = 0.
\]

Proof. Note that since, for any \( x_n \to x \) in \( X \), we have

\[
\lim_{n \to \infty} \sup_{\lambda \in \Lambda} |f_\lambda(x_n) - f_\lambda(x)| = 0, \quad (4.9)
\]

we see that \( \{f_\lambda\}_{\lambda \in \Lambda} \) is an equicontinuous family of functions. Thus, by the Arzela-Ascoli Theorem \[31\], for any given compact set \( K \subset X \) and \( \epsilon > 0 \), there is a finite set of continuous and bounded functions \( F := \{f_1, \ldots, f_N\} \), so that, for any \( \lambda \in \Lambda \), there is \( f_i \in F \) with

\[
\sup_{x \in K} |f_\lambda(x) - f_i(x)| \leq \epsilon.
\]

Now, we claim that, for the same \( f_i \in F \), we have \( \sup_{x \in K} |f_{n,\lambda}(x) - f_i(x)| \leq 3\epsilon/2 \) for large enough \( n \), which is independent of \( \lambda \). To see this, observe the following:

\[
\sup_{x \in K} |f_{n,\lambda}(x) - f_i(x)| \leq \sup_{x \in K} |f_{n,\lambda}(x) - f_\lambda(x)| + \sup_{x \in K} |f_\lambda(x) - f_i(x)|.
\]

Note that the second term is less than \( \epsilon \) and the first term can be made arbitrarily small as \( f_{n,\lambda} \to f_\lambda \) uniformly on compact sets and on \( \Lambda \), which can be easily proved using the assumptions in the lemma.

Note that \( \mu_n \to \mu \) weakly. Hence, \( \{\mu_n\} \) is a tight family of probability measures by Prokhorov theorem \[14\] Theorem 5.2]. Therefore, for any \( \epsilon > 0 \), there exists a compact
subset $K_\epsilon$ of $X$ such that, for all $n$,

$$\mu_n(K_\epsilon) \geq 1 - \epsilon.$$ 

Now, we fix any $\epsilon > 0$ and choose a compact set $K_\epsilon$ such that, for all $n$, $\mu_n(K_\epsilon) \geq 1 - \epsilon$. We also fix a finite family of continuous and bounded functions $F := \{f_1, \ldots, f_N\}$ such that, for any $\lambda$, we can find $f_i \in F$ with $\sup_{x \in K_\epsilon} |f_\lambda(x) - f_i(x)| \leq \epsilon$. Moreover, we choose a large $N$ such that $\sup_{x \in K_\epsilon} |f_n,\gamma(x) - f_i(x)| \leq 3\epsilon/2$ for all $n \geq N$.

With this setup, we go back to the main statement:

\[
\begin{aligned}
\sup_{\lambda \in \Lambda} \left| \int f_{n,\lambda}(x) \mu_n(dx) - \int f_\lambda(x) \mu(dx) \right| \\
\leq \sup_{\lambda \in \Lambda} \left| \int_{X \setminus K_\epsilon} f_{n,\lambda}(x) \mu_n(dx) - \int_{X \setminus K_\epsilon} f_\lambda(x) \mu(dx) \right| \\
+ \sup_{\lambda \in \Lambda} \left| \int_{K_\epsilon} f_{n,\lambda}(x) \mu_n(dx) - \int_{K_\epsilon} f_\lambda(x) \mu(dx) \right| \\
\leq 2\epsilon C + \sup_{\lambda \in \Lambda} \left| \int_{K_\epsilon} (f_{n,\lambda}(x) - f_i(x)) \mu_n(dx) + \int_{K_\epsilon} f_i(x) \mu_n(dx) - \int_{K_\epsilon} f_i(x) \mu(dx) \\
+ \int_{K_\epsilon} (f_i(x) - f_\lambda(x)) \mu(dx) \right| \\
\leq 2\epsilon C + \left| \int_{K_\epsilon} f_i(x) \mu_n(dx) - \int_{K_\epsilon} f_i(x) \mu(dx) \right| + 5\epsilon/2 \\
\leq 2\epsilon C + \left| \int_X f_i(x) \mu_n(dx) - \int_X f_i(x) \mu(dx) \right| \\
+ \left| \int_{K_\epsilon} f_i(x) \mu(dx) - \int_{K_\epsilon} f_i(x) \mu_n(dx) \right| + 5\epsilon/2 \\
\leq 4\epsilon C + 5\epsilon/2 + \left| \int_X f_i(x) \mu_n(dx) - \int_X f_i(x) \mu(dx) \right|
\end{aligned}
\]

where $C$ is the uniform bound on $\{f_{n,\lambda}\}$ and $\{f_\lambda\}$. Since $\epsilon$ is arbitrary and $\mu_n$ converges
weakly to $\mu$, the result follows.

In Theorem 4.4.1 and Theorem 4.4.2, we need to show that, for every $(z_0^n, u_n) \to (z_0, u)$ in $Z \times U$, we have

$$\sup \|f\|_{BL} \leq 1 \left| \int_Z f(z_1) \eta(dz_1|z_0^n, u_n) - \int_Z f(z_1) \eta(dz_1|z_0, u) \right| \to 0,$$

where we equip $Z$ with the metric $\rho$ to define bounded-Lipschitz norm $\|f\|_{BL}$ of any Borel measurable function $f : Z \to \mathbb{R}$. We can equivalently write this as

$$\sup \|f\|_{BL} \leq 1 \left| \int_Y f(z_1(z_0^n, u_n, y_1)) P(dy_1|z_0^n, u_n) - \int_Y f(z_1(z_0, u, y_1)) P(dy_1|z_0, u) \right| \to 0.
\tag{4.10}$$

The term in equation (4.10) can be upper bounded as follows:

$$\sup \|f\|_{BL} \leq 1 \left| \int_Y f(z_1(z_0^n, u_n, y_1)) P(dy_1|z_0^n, u_n) - \int_Y f(z_1(z_0, u, y_1)) P(dy_1|z_0, u) \right|$$

$$\leq \sup \|f\|_{BL} \leq 1 \left| \int_Y f(z_1(z_0^n, u_n, y_1)) P(dy_1|z_0^n, u_n) - \int_Y f(z_1(z_0, u, y_1)) P(dy_1|z_0, u) \right|$$

$$+ \sup \|f\|_{BL} \leq 1 \int_Y \left| f(z_1(z_0^n, u_n, y_1)) - f(z_1(z_0, u, y_1)) \right| P(dy_1|z_0, u)$$

$$\leq \|P(\cdot|z_0^n, u_n) - P(\cdot|z_0, u)\|_{TV}$$

$$+ \sup \|f\|_{BL} \leq 1 \int_Y \left| f(z_1(z_0^n, u_n, y_1)) - f(z_1(z_0, u, y_1)) \right| P(dy_1|z_0, u), \tag{4.11}$$

where, in the last inequality, we have used $\|f\|_{\infty} \leq \|f\|_{BL} \leq 1$. To prove that (4.11) (and so (4.10)) goes to 0, it is sufficient to establish the following results:

(i) $P(dy_1|z_0, u_0)$ is continuous in total variation,
4.4. MAIN RESULTS AND CONNECTIONS WITH THE LITERATURE

(ii) \( \lim_{n \to \infty} \int_Y \rho(z_1^n(z_0^n, u_n, y_1), z_1(z_0, u, y_1)) P(dy_1|z_0, u) = 0 \) as \((z_0^n, u_n) \to (z_0, u)\).

Indeed, suppose that (i) and (ii) hold. Then, the first term in (4.11) goes to 0 as \(P(\cdot|z_0, u)\) is continuous in total variation. For the second term in (4.11), we have

\[
\sup_{\|f\|_{BL} \leq 1} \int_Y \left| f(z_1^n(z_0^n, u_n, y_1)) - f(z_1(z_0, u, y_1)) \right| P(dy_1|z_0, u) \\
\leq \int_Y \rho(z_1^n(z_0^n, u_n, y_1), z_1(z_0, u, y_1)) P(dy_1|z_0, u) \\
\to 0 \text{ as } n \to \infty \quad \text{(by (ii))}.
\]

Therefore, to complete the proof of Theorem 4.4.1 and Theorem 4.4.2, we will prove (i) and (ii).

**Proof of Theorem 4.4.1**

We first prove (i); that is, \(P(dy_1|z_0, u)\) is continuous in total variation. To this end, let \((z_0^n, u_n) \to (z_0, u)\). Then, we write

\[
\sup_{A \in \mathcal{B}(Y)} \left| P(A|z_0^n, u_n) - P(A|z_0, u) \right| \\
= \sup_{A \in \mathcal{B}(Y)} \left| \int_X Q(A|x_1, u_n) T(dx_1|z_0^n, u_n) - \int_X Q(A|x_1, u) T(dx_1|z_0, u) \right|,
\]

where \(T(dx_1|z_0^n, u_n) : = \int_X T(dx_1|x_0, u_n) z_0^n(dx_0)\). Note that, by Lemma 4.4.1, we can show that \(T(dx_1|z_0^n, u_n) \to T(dx_1|z_0, u)\) weakly. Indeed, if \(g \in C_b(X)\), then we define

\[
r_n(x_0) = \int_X g(x_1) T(dx_1|x_0, u_n) \text{ and } r(x_0) = \int_X g(x_1) T(dx_1|x_0, u).
\]

Since \(T(dx_1|x_0, u)\) is weakly continuous, we have \(r_n(x_0^n) \to r(x_0)\) when \(x_0^n \to x_0\). Hence, by Lemma 4.4.1,
we have
\[
\lim_{n \to \infty} \left| \int_X r_n(x_0) z_0^n (dx_0) - \int_X r(x_0) z_0 (dx_0) \right| = 0.
\]

Hence, \(T(dx_1|z_0^n, u_n) \to T(dx_1|z_0, u)\) weakly. Moreover, the families of functions \(\{Q(A|\cdot, u_n)\}_{n \geq 1, A \in B(Y)}\) and \(\{Q(A|\cdot, u)\}_{A \in B(Y)}\) satisfy the conditions of Lemma 4.4.1 as \(Q\) is continuous in total variation distance. Therefore, Lemma 4.4.1 yields that
\[
\lim_{n \to \infty} \sup_{A \in B(Y)} \left| \int_X Q(A|x_1, u_n) T(dx_1|z_0^n, u_n) - \int_X Q(A|x_1, u) T(dx_1|z_0, u) \right| = 0.
\]

Thus, \(P(dy_1|z_0, u)\) is continuous in total variation.

To prove (ii), we write
\[
\int_Y \rho(z_1(z_0^n, u_n, y_1), z_1(z_0, u, y_1)) P(dy_1|z_0, u)
\]
\[
= \int_Y \sum_{m=1}^{\infty} 2^{-m+1} \left| \int_X f_m(x_1) z_1(z_0^n, u_n, y_1)(dx_1) - \int_X f_m(x_1) z_1(z_0, u, y_1)(dx_1) \right| P(dy_1|z_0, u)
\]
\[
= \sum_{m=1}^{\infty} 2^{-m+1} \left| \int_X f_m(x_1) z_1(z_0^n, u_n, y_1)(dx_1) - \int_X f_m(x_1) z_1(z_0, u, y_1)(dx_1) \right| P(dy_1|z_0, u),
\]
where we have used Fubini’s theorem with the fact that \(\sup_m \|f_m\|_{\infty} \leq 1\). For each \(m\), let us define
\[
I_+^{(n)} := \left\{ y_1 \in Y : \int_X f_m(x_1) z_1(z_0^n, u_n, y_1)(dx_1) > \int_X f_m(x_1) z_1(z_0, u, y_1)(dx_1) \right\}
\]
Then, we can write

\[
I_{-}^{(n)} := \left\{ y_{1} \in \mathbb{V} : \int_{\mathcal{X}} f_{m}(x_{1})z_{1}(z_{0}^{n}, u_{n}, y_{1})(dx_{1}) \leq \int_{\mathcal{X}} f_{m}(x_{1})z_{1}(z_{0}, u, y_{1})(dx_{1}) \right\}.
\]

(4.12)

In the sequel, we only consider the term with the set \(I_{+}^{(n)}\). The analysis for the other one follows from the same steps. We have

\[
\int_{\mathcal{X}} \int_{\mathcal{X}} f_{m}(x_{1})z_{1}(z_{0}^{n}, u_{n}, y_{1})(dx_{1}) - \int_{\mathcal{X}} f_{m}(x_{1})z_{1}(z_{0}, u, y_{1})(dx_{1}) \right| P(dy_{1}|z_{0}, u)
\]

\[
= \int_{I_{+}^{(n)}} \left( \int_{\mathcal{X}} f_{m}(x_{1})z_{1}(z_{0}^{n}, u_{n}, y_{1})(dx_{1}) - \int_{\mathcal{X}} f_{m}(x_{1})z_{1}(z_{0}, u, y_{1})(dx_{1}) \right) P(dy_{1}|z_{0}, u)
\]

\[
+ \int_{I_{-}^{(n)}} \left( \int_{\mathcal{X}} f_{m}(x_{1})z_{1}(z_{0}, u, y_{1})(dx_{1}) - \int_{\mathcal{X}} f_{m}(x_{1})z_{1}(z_{0}^{n}, u_{n}, y_{1})(dx_{1}) \right) P(dy_{1}|z_{0}, u).
\]
where we have used $\|f_m\|_\infty \leq 1$ in the last inequality. The first term above goes to 0 since $P(dy_1|z_0, u)$ is continuous in total variation. For the second term, we use Lemma 4.4.1. Indeed, families of functions $\{f_m(\cdot)Q(A|\cdot, u_n) : n \geq 1, A \in \mathcal{B}(Y)\}$ and $\{f_m(\cdot)Q(A|\cdot, u) : A \in \mathcal{B}(Y)\}$ satisfy the conditions in Lemma 4.4.1 as $Q$ is continuous in total variation. Hence, the second term converges to 0 by Lemma 4.4.1 since $\mathcal{T}(dx_1|z_0^n, u_n) \rightarrow \mathcal{T}(dx_1|z_0, u)$ weakly. Hence, for each $m$, we have

$$\lim_{n \to \infty} \int_Y \left| \int_X f_m(x_1)z_1(z_0^n, u_n, y_1)(dx_1) \right| P(dy_1|z_0, u) = 0.$$ 

By the dominated convergence theorem, we then have

$$\lim_{n \to \infty} \int_Y \rho(z_1(z_0^n, u_n, y_1), z_1(z_0, u, y_1))P(dy_1|z_0, u)$$

$$\leq \sum_{m=1}^{\infty} 2^{-m+1} \lim_{n \to \infty} \int_Y \left| \int_X f_m(x_1)z_1(z_0^n, u_n, y_1)(dx_1) \right| P(dy_1|z_0, u) = 0.$$ 

This establishes (ii), which completes the proof together with (i).

**Proof of Theorem 4.4.2**

We first show (i); that is, $P(dy_1|z_0, u_0)$ is continuous total variation. Let $(z_0^n, u_n) \rightarrow (z_0, u)$. Then, we have

$$\sup_{A \in \mathcal{B}(Y)} |P(A|z_0^n, u_n) - P(A|z_0, u)|$$
4.4. MAIN RESULTS AND CONNECTIONS WITH THE LITERATURE

\[
\sup_{A \in \mathcal{B}(\mathcal{Y})} \left| \int_{\mathcal{X}} \int_{\mathcal{X}} Q(A|x_1) T(dx_1|x_0, u_n) z_0^n (dx_0) - \int_{\mathcal{X}} \int_{\mathcal{X}} Q(A|x_1) T(dx_1|x_0, u) z_0 (dx_0) \right|
\]

For each \(A \in \mathcal{B}(\mathcal{Y})\) and \(n \geq 1\), we define

\[
f_{n,A}(x_0) = \int_{\mathcal{X}} Q(A|x_1) T(dx_1|x_0, u_n)
\]

and

\[
f_A(x_0) = \int_{\mathcal{X}} Q(A|x_1) T(dx_1|x_0, u).
\]

Then, for all \(x_0^n \to x_0\), we have

\[
\lim_{n \to \infty} \sup_{A \in \mathcal{B}(\mathcal{Y})} |f_{n,A}(x_0^n) - f_A(x_0)|
\]

\[
= \lim_{n \to \infty} \sup_{A \in \mathcal{B}(\mathcal{Y})} \left| \int_{\mathcal{X}} Q(A|x_1) T(dx_1|x_0^n, u_n) - \int_{\mathcal{X}} Q(A|x_1) T(dx_1|x_0, u) \right|
\]

\[
\leq \lim_{n \to \infty} \|T(dx_1|x_0^n, u_n) - T(dx_1|x_0, u)\|_{TV} = 0
\]

and

\[
\lim_{n \to \infty} \sup_{A \in \mathcal{B}(\mathcal{Y})} |f_A(x_0^n) - f_A(x_0)|
\]

\[
= \lim_{n \to \infty} \sup_{A \in \mathcal{B}(\mathcal{Y})} \left| \int_{\mathcal{X}} Q(A|x_1) T(dx_1|x_0^n, u) - \int_{\mathcal{X}} Q(A|x_1) T(dx_1|x_0, u) \right|
\]

\[
\leq \lim_{n \to \infty} \|T(dx_1|x_0^n, u) - T(dx_1|x_0, u)\|_{TV} = 0.
\]
Then, by Lemma [4.4.1] we have

\[
\lim_{n \to \infty} \sup_{A \in \mathcal{B}(Y)} \left| \int_X f_{n,A}(x_0) z_0^n(dx_0) - \int_X f_A(x_0) z_0(dx_0) \right|
\]

\[
= \lim_{n \to \infty} \sup_{A \in \mathcal{B}(Y)} \left| \int_X \int_X Q(A|x_1) \mathcal{T}(dx_1|x_0, u_n) z_0^n(dx_0) \right|
- \left| \int_X \int_X Q(A|x_1) \mathcal{T}(dx_1|x_0, u) z_0(dx_0) \right|
\]

\[
= 0.
\]

Hence, \( P(dy_1|z_0, u_0) \) is continuous in total variation.

Now, we show (ii); that is, for any \((z_0^n, u_n) \to (z_0, u)\), we have

\[
\lim_{n \to \infty} \int_Y \rho(z_1(z_0^n, u_n, y_1), z_1(z_0, u, y_1)) P(dy_1|z_0, u) = 0.
\]

From the proof of Theorem [4.4.1] it suffices to show that

\[
\lim_{n \to \infty} \int_X \int_{I_+^{(n)}} f_m(x_1) Q(dy_1|x_1) \mathcal{T}(dx_1|z_0^n, u_n)
- \int_X \int_{I_+^{(n)}} f_m(x_1) Q(dy_1|x_1) \mathcal{T}(dx_1|z_0, u) = 0. \tag{4.13}
\]

Indeed, we have

\[
\left| \int_X \int_{I_+^{(n)}} f_m(x_1) Q(dy_1|x_1) \mathcal{T}(dx_1|z_0^n, u_n) - \int_X \int_{I_+^{(n)}} f_m(x_1) Q(dy_1|x_1) \mathcal{T}(dx_1|z_0, u) \right|
\]

\[
\leq \left| \int_X f_m(x_1) Q(I_+^{(n)}|x_1) \mathcal{T}(dx_1|x_0, u_n) z_0^n(dx_0) \right|
- \left| \int_X f_m(x_1) Q(I_+^{(n)}|x_1) \mathcal{T}(dx_1|x_0, u) z_0^n(dx_0) \right|
\]
4.4. MAIN RESULTS AND CONNECTIONS WITH THE LITERATURE

\[
\begin{align*}
+ & \left| \int_{X^2} f_m(x_1) Q(I_+^{(n)} | x_1) \mathcal{T}(dx_1 | x_0, u) z_0^n(dx_0) \right| \\
- & \left| \int_{X^2} f_m(x_1) Q(I_+^{(n)} | x_1) \mathcal{T}(dx_1 | x_0, u) z_0(dx_0) \right|
\end{align*}
\]

\[
\leq \int_X \| \mathcal{T}(dx_1 | x_0, u_n) - \mathcal{T}(dx_1 | x_0, u) \|_{TV} z_0^n(dx_0)
\]

\[
+ \left| \int_{X^2} f_m(x_1) Q(I_+^{(n)} | x_1) \mathcal{T}(dx_1 | x_0, u) z_0^n(dx_0) \right| \\
- \left| \int_{X^2} f_m(x_1) Q(I_+^{(n)} | x_1) \mathcal{T}(dx_1 | x_0, u) z_0(dx_0) \right|
\]

where we have used \( \sup_{n \geq 1} \sup_{x_1 \in X} |f_m(x_1) Q(I_+^{(n)} | x_1)| \leq 1 \) in the last inequality. If we define \( r_n(x_0) = \| \mathcal{T}(dx_1 | x_0, u_n) - \mathcal{T}(dx_1 | x_0, u) \|_{TV} z_0^n(dx_0) \), then \( r_n(x_0^n) \to 0 \) whenever \( x_0^n \to x_0 \).

Then, the first term converges to 0 by Lemma 4.4.1 as \( z_0^n \to z_0 \) weakly. The second term also converges to 0 by Lemma 4.4.1 since \( \{ \int_X f(x_1) Q(I_+^{(n)} | x_1) \mathcal{T}(dx_1 | x_0, u) : n \geq 1 \} \) is a family of uniformly bounded and equicontinuous functions by total variation continuity of \( \mathcal{T}(dx_1 | x_0, u) \). This proves (ii) and completes the proof together with (i).

Proof of Theorem 4.4.3

Let \( S \) be a separable metric space. Another metric that metrizes the weak topology on \( \mathcal{P}(S) \) is the following:

\[
\rho(\mu, \nu) = \sum_{m=1}^{\infty} 2^{-(m+1)} \left| \int_S f_m(x) \mu(dx) - \int_S f_m(x) \nu(dx) \right|
\]

where \( \{ f_m \}_{m \geq 1} \) is an appropriate sequence of continuous and bounded functions such that \( \| f_m \|_\infty \leq 1 \) for all \( m \geq 1 \) (see [82, Theorem 6.6, p. 47]).

We equip \( Z \) with the metric \( \rho \) to define bounded-Lipschitz norm \( \| f \|_{BL} \) of any Borel
measurable function \( f : Z \to \mathbb{R} \). With this metric, we can start the proof:

\[
\sup_{\|f\|_{BL} \leq 1} \left| \int_Z f(z_1) \eta(dz_1|z_0', u) - \int_Z f(z_1) \eta(dz_1|z_0, u) \right|
\]

\[
= \sup_{\|f\|_{BL} \leq 1} \left| \int_Y f(z_0'|u,y_1) P(dy_1|z_0', u) - \int_Y f(z_0|u,y_1) P(dy_1|z_0, u) \right|
\]

\[
\leq \sup_{\|f\|_{BL} \leq 1} \left| \int_Y f(z_0'|u,y_1) P(dy_1|z_0', u) - \int_Y f(z_0'|u,y_1) P(dy_1|z_0, u) \right|
\]

\[
+ \sup_{\|f\|_{BL} \leq 1} \left| \int_Y f(z_0'|u,y_1) - f(z_0|u,y_1) \right| P(dy_1|z_0, u)
\]

\[
\leq \|P(\cdot|z_0', u) - P(\cdot|z_0, u)\|_{TV}
\]

\[
+ \sup_{\|f\|_{BL} \leq 1} \left| \int_Y f(z_0'|u,y_1) - f(z_0|u,y_1) \right| P(dy_1|z_0, u)
\]

where, in the last inequality, we have used \( \|f\|_{\infty} \leq \|f\|_{BL} \leq 1 \).

We first analyze the first term:

\[
\|P(\cdot|z_0', u) - P(\cdot|z_0, u)\|_{TV} = \sup_{A \in \mathcal{B}(Y)} |P(A|z_0', u) - P(A|z_0, u)|
\]

\[
= \sup_{A \in \mathcal{B}(Y)} \left| \int_X \int_X Q(A|x_1) \mathcal{T}(dx_1|x_0, u) z_0'(dx_0) \right|
\]

\[
- \int_X \int_X Q(A|x_1) \mathcal{T}(dx_1|x_0, u) z_0(dx_0)
\]

For each \( A \in \mathcal{B}(Y) \) we define

\[
f_A(x_0) = \int_X Q(A|x_1) \mathcal{T}(dx_1|x_0, u).
\]

Then, for all \( x_0', x_0 \) and for any \( A \in \mathcal{B}(Y) \), we have

\[
|f_A(x_0') - f_A(x_0)| = \left| \int_X Q(A|x_1) \mathcal{T}(dx_1|x_0', u) - \int_X Q(A|x_1) \mathcal{T}(dx_1|x_0, u) \right|
\]
4.4. MAIN RESULTS AND CONNECTIONS WITH THE LITERATURE

\[ \leq \| \mathcal{T}(\cdot|x_0', u) - \mathcal{T}(\cdot|x_0, u) \|_{TV} \leq \alpha_X |x'_0 - x_0|. \]

Hence, we have that \( \| f_A \|_{BL} \leq 1 + \alpha_X \). Then, we can write

\[ \| P(\cdot|z'_0, u) - P(\cdot|z_0, u) \|_{TV} \leq \sup_{A \in \mathcal{B}(Y)} \left| \int_X f_A(x_0) z'_0(dx_0) - \int_X f_A(x_0) z_0(dx_0) \right| \]

\[ \leq (1 + \alpha_X) \rho_{BL}(z'_0, z_0). \]

Now we focus on term (4.15):

\[
\sup_{\|f\|_{BL} \leq 1} \int_Y |f(z_1(z'_0, u, y_1)) - f(z_1(z_0, u, y_1))| P(dy_1|z_0, u) \\
\leq \int_Y \rho(z_1(z'_0, u, y_1), z_1(z_0, u, y_1)) P(dy_1|z_0, u) \\
= \int_Y \sum_{m=1}^{\infty} 2^{-m+1} \left| \int_X f_m(x_1) z_1(z'_0, u, y_1)(dx_1) - \int_X f_m(x_1) z_1(z_0, u, y_1)(dx_1) \right| P(dy_1|z_0, u) \\
= \sum_{m=1}^{\infty} 2^{-m+1} \int_Y \left| \int_X f_m(x_1) z_1(z'_0, u, y_1)(dx_1) - \int_X f_m(x_1) z_1(z_0, u, y_1)(dx_1) \right| P(dy_1|z_0, u),
\]

where we have used Fubini’s theorem with the fact that \( \sup_m \| f_m \|_{\infty} \leq 1 \). For each \( m \), let us define

\[ I_+ := \left\{ y_1 \in Y : \int_X f_m(x_1) z_1(z'_0, u, y_1)(dx_1) > \int_X f_m(x_1) z_1(z_0, u, y_1)(dx_1) \right\} \]
\[ I_- := \left\{ y_1 \in Y : \int_X f_m(x_1) z_1(z'_0, u, y_1)(dx_1) \leq \int_X f_m(x_1) z_1(z_0, u, y_1)(dx_1) \right\}. \]
Then, we can write

\[
\int_X \left( \int_X f_m(x_1)z_1(z_0', u, y_1)(dx_1) - \int_X f_m(x_1)z_1(z_0, u, y_1)(dx_1) \right) P(dy_1|z_0, u)
= \int_{I_+} \left( \int_X f_m(x_1)z_1(z_0', u, y_1)(dx_1) - \int_X f_m(x_1)z_1(z_0, u, y_1)(dx_1) \right) P(dy_1|z_0, u)
+ \int_{I_-} \left( \int_X f_m(x_1)z_1(z_0, u, y_1)(dx_1) - \int_X f_m(x_1)z_1(z_0', u, y_1)(dx_1) \right) P(dy_1|z_0, u).
\]

In the sequel, we only consider the term with the set \( I_+ \). The analysis for the other one follows from the same steps. We have

\[
\int_{I_+} \left( \int_X f_m(x_1)z_1(z_0', u, y_1)(dx_1) - \int_X f_m(x_1)z_1(z_0, u, y_1)(dx_1) \right) P(dy_1|z_0, u)
\leq \int_{I_+} \int_X f_m(x_1)z_1(z_0', u, y_1)(dx_1) P(dy_1|z_0, u)
- \int_{I_+} \int_X f_m(x_1)z_1(z_0', u, y_1)(dx_1) P(dy_1|z_0', u)
+ \int_{I_+} \int_X f_m(x_1)z_1(z_0, u, y_1)(dx_1) P(dy_1|z_0, u)
- \int_{I_+} \int_X f_m(x_1)z_1(z_0, u, y_1)(dx_1) P(dy_1|z_0, u)
\leq \| P(dy_1|z_0, u) - P(dy_1|z_0', u) \|_{TV}
+ \int_X \int_{I_+} f_m(x_1)Q(dy_1|x_1)T(dx_1|z_0', u) - \int_X \int_{I_+} f_m(x_1)Q(dy_1|x_1)T(dx_1|z_0, u)
\leq (1 + \alpha_X) \rho_{BL}(z_0', z_0)
+ \int_X \int_{I_+} f_m(x_1)Q(dy_1|x_1)T(dx_1|z_0', u) - \int_X \int_{I_+} f_m(x_1)Q(dy_1|x_1)T(dx_1|z_0, u),
\]

where we have used \( \| f_m \|_\infty \leq 1 \) in the last inequality. For the last term, we write

\[
\int_X \int_{I_+} f_m(x_1)Q(dy_1|x_1)T(dx_1|z_0', u) - \int_X \int_{I_+} f_m(x_1)Q(dy_1|x_1)T(dx_1|z_0, u)
\]
For the last step, we use the fact, for \( \alpha_1 + g \leq \infty \), that

\[
\|T'(\cdot | x'_0, u) - T'(\cdot | x_0, u)\|_{TV} \leq \alpha_X |x'_0 - x_0|.
\]

Notice that, the same steps can be used for \( I_- \) to get the same bound. Hence, we can write that

\[
\begin{align*}
\int_{Y} \left| \int_{X} f_m(x_1) z_1(z'_0, u, y_1)(dx_1) - \int_{X} f_m(x_1) z_1(z_0, u, y_1)(dx_1) \right| P(dy_1 | z_0, u) \\
= \int_{I_+} \left( \int_{X} f_m(x_1) z_1(z'_0, u, y_1)(dx_1) - \int_{X} f_m(x_1) z_1(z_0, u, y_1)(dx_1) \right) P(dy_1 | z_0, u) \\
+ \int_{I_-} \left( \int_{X} f_m(x_1) z_1(z_0, u, y_1)(dx_1) - \int_{X} f_m(x_1) z_1(z'_0, u, y_1)(dx_1) \right) P(dy_1 | z_0, u) \\
\leq 4(1 + \alpha_X) \rho_{BL}(z'_0, z_0)
\end{align*}
\]

Now we go back to the term (4.13):

\[
\begin{align*}
\sup_{\|f\|_{BL} \leq 1} & \int_{Y} \left| f(z_1(z'_0, u, y_1)) - f(z_1(z_0, u, y_1)) \right| P(dy_1 | z_0, u) \\
\leq & \int_{Y} \rho(z_1(z'_0, u, y_1), z_1(z_0, u, y_1)) P(dy_1 | z_0, u) \\
= & \int_{Y} \sum_{m=1}^{\infty} 2^{-m+1} \left| \int_{X} f_m(x_1) z_1(z'_0, u, y_1)(dx_1) - \int_{X} f_m(x_1) z_1(z_0, u, y_1)(dx_1) \right| P(dy_1 | z_0, u)
\end{align*}
\]
4.4. MAIN RESULTS AND CONNECTIONS WITH THE LITERATURE

\[
= \sum_{m=1}^{\infty} 2^{-m+1} \int_{\mathcal{X}} \int_{\mathcal{Y}} f_m(x_1) z_1(x', u, y_1) (dx_1) - \int_{\mathcal{X}} f_m(x_1) z_1(x, u, y_1) (dx_1) \left| P(dy_1 | z_0, u) \right|
\leq 4(1 + \alpha_X) \rho_{BL}(z'_0, z_0).
\]

Thus, combining all the results:

\[
\sup_{\|f\|_{BL} \leq 1} \left| \int_{\mathcal{Z}} f(z_1) \eta(dz_1 | z'_0, u) - \int_{\mathcal{Z}} f(z_1) \eta(dz_1 | z_0, u) \right|
\leq \|P(\cdot | z'_0, u) - P(\cdot | z_0, u)\|_{TV}
+ \sup_{\|f\|_{BL} \leq 1} \int_{\mathcal{Y}} \left| f(z_1(x', u, y_1)) - f(z_1(x, u, y_1)) \right| P(dy_1 | z_0, u)
\leq 5(1 + \alpha_X) \rho_{BL}(z'_0, z_0).
\]

4.4.5 A Technical Generalization

In this section, we prove the weak Feller property of the filter process under more general condition than those in Theorem 4.4.1 and Theorem 4.4.2. But, we note that it is indeed generally infeasible to establish this condition without imposing assumptions similar to the assumptions in Theorem 4.4.1 and Theorem 4.4.2. Therefore, although this condition is more general than those in Theorem 4.4.1 and Theorem 4.4.2, this generalization is not excessively important in practice.

We first note that our proof technique brings to light the main ingredients that is necessary to prove the weak Feller property of the filter process via the item (i) and eq. (4.13); that is,

- \(P(dy_1 | z_0, u_0)\) is continuous in total variation,
- \(\lim_{n \to \infty} \int_{\mathcal{X}} f_m(x_1) Q(I_+^{(n)} | x_1) \mathcal{T}(dx_1 | z_0^n, u_n) - \int_{\mathcal{X}} f_m(x_1) Q(I_+^{(n)} | x_1) \mathcal{T}(dx_1 | z_0, u) = 0.\)
This observation suggests the following condition that generalize the conditions in our previously stated main results. Let $\mathcal{F} = \{ f_m \}_{m \geq 1} \subset C_b(\mathcal{X})$ be a countable set of continuous and bounded functions such that $\|f_m\|_\infty \leq 1$ for all $m \geq 1$, $1_\mathcal{X} \in \mathcal{F}$, and $\mathcal{F}$ metrizes the weak topology on $\mathcal{P}(\mathcal{X})$ via the metric $\rho$ introduced in Section 4.2. Then, we state the following assumption:

(M) For each $f \in \mathcal{F}$, the family of functions

$$(z_0, u_0) \mapsto \int_{\mathcal{X}} f(x_1)Q(A|x_1, u_0)T(dx_1|z_0, u_0)$$

is equicontinuous when indexed by $A \in \mathcal{B}(\mathcal{Y})$.

Using Lemma 4.4.1, it is fairly straightforward to prove that conditions in Theorem 4.4.1 and Theorem 4.4.2 both imply the assumption (M). Hence, assumption (M) is more general than those in Theorem 4.4.1 and Theorem 4.4.2.

**Theorem 4.4.5.** Under assumption (M), the transition probability $\eta(\cdot|z, u)$ of the filter process is weakly continuous in $(z, u)$.

**Proof.** Recall that it is sufficient to prove the following:

(i) $P(dy_1|z_0, u_0)$ is continuous in total variation,

(ii) $\lim_{n \to \infty} \int_{\mathcal{X}} \rho(z_1(z_0^n, u_n, y_1), z_1(z_0, u, y_1))P(dy_1|z_0, u) = 0$ as $(z_0^n, u_n) \to (z_0, u)$.

Firstly, (i) is true since $1_\mathcal{X} \in \mathcal{F}$. For (ii), it suffices to show that

$$\lim_{n \to \infty} \int_{\mathcal{X}} f_m(x_1)Q(I_+^{(n)}|x_1)T(dx_1|z_0^n, u_n) - \int_{\mathcal{X}} f_m(x_1)Q(I_+^{(n)}|x_1)T(dx_1|z_0, u) = 0.$$

But this immediately follows from assumption (M).
A careful look at the proof of the weak Feller property of the filter process in Feinberg et. al. [42] reveals that they have first established the weak Feller property under a condition somewhat similar to the assumption (M), and then, establish Theorem 4.4.1 by proving that assumptions in Theorem 4.4.1 imply this more general condition. Indeed, let $\tau_b = \{O_j\} \subset X$ be a countable base for the topology on $X$ such that $X \in \tau_b$. Then, under the following assumption:

**(F)** For each finite intersection $O = \bigcap_{n=1}^N O_{j_n}$, where $O_{j_n} \in \tau_b$, the family of functions

$$(z_0, u_0) \mapsto \int_X \int_X 1_{O}(x_1)Q(A|x_1, u_0)T(dx_1|x_0, u_0)z_0(dx_0)$$

is equicontinuous when indexed by $A \in \mathcal{B}(\mathcal{Y})$,

they have proved that the weak Feller property of the filter process holds (see [42, Lemma 5.3] and [40, Theorem 5.5]). We observe that (F) is very similar to (M) except that, in (F), Feinberg et. al. use open sets instead of continuous and bounded functions. However, proving that conditions in Theorem 4.4.1 imply the assumption (F) as in [42] requires quite tedious mathematical methods. By using open sets instead of continuous and bounded functions, one needs to work with inequalities and limit infimum operation as a result of Portmanteau theorem [4, Theorem 2.1] (and the associated proof program involving generalized Fatou’s lemma [39] in place of equalities and limit operation, which are significantly easier to analyze than the former leading to a much more concise analysis that we have presented in this section. For instance, [42, Theorem 5.1] is the key result to prove that the weak Feller condition of the transition probability and the total variation continuity of the observation channel imply the assumption (F). We note that if one states this result using continuous and bounded functions in place of open sets, then this version of [42, Theorem
5.1] becomes a corollary of Lemma 4.4.1, which has a concise and easy to follow proof. But, the proof of [42, Theorem 5.1] with open sets requires quite tedious mathematical concepts from topology and weak convergence of probability measures. In view of this discussion, we also note that Theorem 4.4.2 can also be proved using the condition (F) rather than our approach building on (M) through some additional argumentation.

In summary, our approach allows for a more direct and concise approach which also makes the proof of Theorem 4.4.1 more accessible. Once again, we note that Theorem 4.4.2 has not been reported in the literature.

4.5 Conclusion

In this chapter, there are two main contributions: (i) the weak Feller property of the filter process is established under a new condition, which assumes that the state transition probability is continuous under the total variation convergence with no assumptions on the measurement model, and (ii) a concise and easy to follow proof of the same result under the weak Feller condition of the transition probability and the total variation continuity of the observation channel, which was first established in [42], is also given. Implications of these results have also been presented.
Chapter 5

Approximations in Partially Observed Models Using Finite Memory

As discussed earlier, partially observed Markov decision processes (POMDPs) offer a practically rich and relevant, and mathematically challenging model. Even in the most basic setup of finite state-action models, the analysis and computation of optimal solutions is complicated. The existence of optimal policies has in general been established via converting, or reducing, the original partially observed stochastic control problem to a fully observed Markov Decision Problem (MDP) with probability measure valued (belief) states, leading to a belief-MDP. However, computing an optimal policy for this fully observed model, and so for the original POMDP, using classical methods (such as dynamic programming, policy iteration, linear programming) is not simple even if the original system has finite state and action spaces, since the state space of the fully observed (reduced) model is always uncountable.

For this chapter, we are interested in the infinite horizon discounted cost,

\[ J_\beta(\mu, T, \gamma) = E_\mu^{T, \gamma} \left[ \sum_{t=0}^{\infty} \beta^t c(X_t, U_t) \right] \]
For partially observed stochastic problems, the optimal policies use all the available information in general. The question we ask is the following one: suppose that instead of using all available history, we construct an approximate model using the finite window information variables

\[ I_t^N = \begin{cases} Y_{[t-N,t]}, U_{[t-N,t-1]} & \text{if } t \geq N, \\ Y_{[0,t]}, U_{[0,t-1]} & \text{if } 0 < t < N, \\ Y_0 & \text{if } t = 0. \end{cases} \]

(5.1)

that is we observe the information variables through a window whose length is \( N \). Suppose, we denote the optimal value function of the approximate model by \( J^N_\beta \) and the approximate policy by \( \gamma^N \).

Under this setup, we will study the following problem.

**Problem** Under suitable conditions, can we find explicit bounds on \( J_\beta^*(\mu, T, \gamma^N) - J_\beta^*(\mu, T) \) in terms of \( N \) and a constructive approximate solution achieving this bound?

We provide two different approximation schemes for this problem in Section 5.3 and Section 5.4. We then give a detailed comparison of two methods in Section 5.5.

5.1 Literature Review

In the MDP theory, various methods have been developed to compute approximately optimal policies by reducing the original problem into a simpler one. A partial list of these
techniques is as follows: approximate dynamic programming, approximate value or policy iteration, simulation-based techniques, neuro-dynamic programming (or reinforcement learning), state aggregation, etc. ([33, 11, 25]). ([93]) investigated finite action and state approximations of fully observed stochastic control problems with general state and action spaces under the discounted cost and average cost optimality criteria where weak continuity conditions were shown to be sufficient for near optimality of finite state approximations. For the discounted cost case, the aforementioned studies showed that optimal policies obtained from these models asymptotically achieve the optimal cost for the original problem under the weak continuity assumption on the controlled transition kernel.

On POMDPs, however, the problem of approximation is significantly more challenging. Most of the studies in the literature are algorithmic and computational contributions. These include ([85]) and ([122]) which develop computational algorithms, utilizing structural convexity/concavity properties of the value function under the discounted cost criterion. ([109]) provides an insightful algorithm which may be regarded as a quantization of the belief space; however, no rigorous convergence results are provided. References ([99]) and ([84]) also present quantization based algorithms for the belief state, where the state, measurement, and the action sets are finite.

For partially observed setups, ([94]), building on ([93]), introduces a rigorous approximation analysis after establishing weak continuity conditions on the transition kernel defining the (belief-MDP) via the non-linear filter ([41, 61]), and shows that finite model approximations obtained through quantization are asymptotically optimal and the control policies obtained from the finite model can be applied to the actual system with asymptotically vanishing error as the number of quantization bins increases. Another rigorous set of studies is ([119]) and ([120]) where the authors provide an explicit quantization method for the
set of probability measures containing the belief states, where the state space is parametrically representable under strong density regularity conditions. The quantization is done through the approximations as measured by Kullback-Leibler divergence (relative entropy) between probability density functions. Further recent studies include ([75]) and ([101]). ([101]) presents a notion of approximate information variable and studies near optimality of policies that satisfies the approximate information state property. In ([75]), a similar problem is analyzed under a decentralized setup. Our explicit approximation results in this chapter will find applications in both of these studies.

We refer the reader to the survey papers ([73, 111, 48]) and the recent book ([67]) for further structural results as well as algorithmic and computational methods for approximating POMDPs. Notably, for POMDPs ([67]) presents structural results on optimal policies under monotonicity conditions of the value function in the belief variable.

With regard to approximations based on finite memory, the following two papers are particularly relevant to our work:

In ([114]), the authors study near optimality of finite window policies for average cost problems where the state, action and observation spaces are finite; under the condition that the liminf and limsup of the average cost are equal and independent of the initial state, the paper establishes the near-optimality of (non-stationary) finite memory policies. Here, a concavity argument building on Feinberg ([38]) (which becomes consequential by the equality assumption) and the finiteness of the state space is crucial. The paper shows that for any given $\epsilon > 0$, there exists an $\epsilon$-optimal finite window policy. However, the authors do not provide a performance bound related to the length of the window, and in fact the proof method builds on convex analysis. Nonetheless, the constant property of the value functions over initial priors is related to unique ergodicity, and thus the stability problem
of non-linear filters, which is a topic of current investigation particularly in the controlled setup.

Reference ([112]) is another related work that studies finite memory approximation techniques for POMDPs with finite state, action, measurements. The POMDP is reduced to a belief MDP and the worst and best case predictors prior to the $N$ most recent information variables are considered to build an approximate belief MDP. The original value function is bounded using these approximate belief MDPs that use only finite memory, where the finiteness of the state space is critically utilized. Furthermore, a loss bound is provided for a suboptimally constructed policy that only uses finite history, where the bound depends on a more specific ergodicity coefficient (which requires restrictive, sample pathwise, contraction properties). In this chapter, we consider more general signal spaces and more relaxed filter stability conditions, and establish explicit rates of convergence results. We also rigorously establish the relation of the loss bound to nonlinear filter stability and state space reduction techniques for MDPs.

5.2 Regularity and Stability Properties of the Belief-MDP

Recall that (see Section 4.1) any POMDP can be viewed as a belief MDP with a transition kernel $\eta$ and a cost function $\tilde{c}$.

Even though, the belief MDP approach provides a strong tool for the analysis of POMDPs, it is usually too complicated computationally. The belief space $\mathcal{Z} = \mathcal{P}(X)$ is always uncountable even when $X$, $Y$ and $U$ are finite. Furthermore, the information variables $I_t$ grows with time and the computation of the belief state $Pr(X_t \in \cdot | I_t)$ can become intractable. Therefore, approximation of the belief-MDP is usually needed.

Before constructing the approximate models, we provide some filter stability results
5.2. REGULARITY AND STABILITY PROPERTIES OF THE BELIEF-MDP

that will help with the error analysis of the approximation techniques we will construct.

5.2.1 Ergodicity and Filter Stability Properties of Partially Observed MDPs

Given a prior \( \mu \in \mathcal{P}(X) \) and a policy \( \gamma \in \Gamma \), we define the filter and predictor for a POMDP in the following.

**Definition 5.2.1.** The one step predictor process is defined as the sequence of conditional probability measures

\[
\pi_{n-1}^{\mu,\gamma}(\cdot) = P^{\mu,\gamma}(X_n \in \cdot | Y_{[0,n-1]}, U_{[0,n-1]} = \gamma_n(Y_{[0,n-1]}, U_{[0,n-2]})) = P^{\mu,\gamma}(X_n \in \cdot | Y_{[0,n-1]}), \quad n \in \mathbb{N}
\]

where \( P^{\mu,\gamma} \) is the probability measure induced by the prior \( \mu \) and the policy \( \gamma \), when \( \mu \) is the probability measure on \( X_0 \).

**Definition 5.2.2.** The filter process is defined as the sequence of conditional probability measures

\[
\pi_n^{\mu,\gamma}(\cdot) = P^{\mu,\gamma}(X_n \in \cdot | Y_{[0,n]}, U_{[0,n-1]} = \gamma_n(Y_{[0,n-1]}, U_{[0,n-2]})) = P^{\mu,\gamma}(X_n \in \cdot | Y_{[0,n]}), \quad n \in \mathbb{N}
\]

where \( P^{\mu,\gamma} \) is the probability measure induced by the prior \( \mu \) and the policy \( \gamma \).

**Definition 5.2.3.** ([29] Equation 1.16) For a kernel operator \( K : S_1 \rightarrow \mathcal{P}(S_2) \) (that is a regular conditional probability from \( S_1 \) to \( S_2 \)) for standard Borel spaces \( S_1, S_2 \), we define the Dobrushin coefficient as:

\[
\delta(K) = \inf \sum_{i=1}^n \min(K(x, A_i), K(y, A_i))
\]

where the infimum is over all \( x, y \in S_1 \) and all partitions \( \{A_i\}_{i=1}^n \) of \( S_2 \).
5.2. REGULARITY AND STABILITY PROPERTIES OF THE BELIEF-MDP

We note that this definition holds for continuous or finite/countable spaces \( S_1 \) and \( S_2 \) and \( 0 \leq \delta(K) \leq 1 \) for any kernel operator.

**Example 5.2.1.** Assume for a finite setup, we have the following stochastic transition matrix

\[
K = \begin{pmatrix}
\frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\
0 & \frac{1}{2} & \frac{1}{2} \\
\frac{2}{3} & 0 & \frac{1}{4}
\end{pmatrix}
\]

The Dobrushin coefficient is the minimum over any two rows where we sum the minimum elements among those rows. For this example, the first and the second rows give \( \frac{2}{3} \), the first and the third rows give \( \frac{7}{12} \) and the second and the third rows give \( \frac{1}{4} \). Then the Dobrushin coefficient is \( \frac{1}{4} \).

Let

\[
\tilde{\delta}(T) := \inf_{u \in U} \delta(T(\cdot|\cdot), u))
\]

**Definition 5.2.4.** For \( X \subset \mathbb{R}^m \) for some \( m \in \mathbb{N} \), and for two probability measures \( \mu, \nu \in \mathcal{P}(X) \), \( \mu \) is said to be absolutely continuous with respect to \( \nu \) if \( \mu(A) = 0 \) for every set \( A \in \mathcal{B}(X) \) for which \( \nu(A) = 0 \). We denote the absolute continuity of \( \mu \) with respect to \( \nu \) by \( \mu \ll \nu \).

**Theorem 5.2.1.** ([78] Theorem 3.3) Assume that for \( \mu, \nu \in \mathcal{P}(X) \), we have \( \mu \ll \nu \). Then we have

\[
E^{\mu, \gamma} \left[ \left\| \pi_{n+1}^{\mu, \gamma} - \pi_{n+1}^{\nu, \gamma} \right\|_{TV} \right] \leq (1 - \tilde{\delta}(T))(2 - \delta(Q))E^{\mu, \gamma} \left[ \left\| \pi_n^{\mu, \gamma} - \pi_n^{\nu, \gamma} \right\|_{TV} \right].
\]
5.3. Method I: Belief-space Quantization Based on Finite-Memory

In particular, defining \( \alpha := (1 - \tilde{\delta}(T))(2 - \delta(Q)) \), we have

\[
E^{\mu,\gamma} \left[ \| \pi_{n}^{\mu,\gamma} - \pi_{n}^{\nu,\gamma} \|_{TV} \right] \leq 2\alpha^n.
\]

This result will be a key ingredient for our main results. It provides conditions on when the belief-state processes for a given POMDP under different priors get closer when they are fed with same observation processes, in expectation under the true probability space. In a vague sense, if the state process is tracked only using a finite window of recent measurement and control variables (and forgets the past observations and actions), then the amount of mismatch from the true filter can be bounded with an error that is exponentially diminishing with the window size.

5.3 Method I: Belief-space Quantization based on Finite-Memory

In this section, we will introduce our first approximation method for belief MDPs. The construction of the finite model is done by quantizing the original belief state space with reachable posterior distributions using only a finite memory of history and a fixed initial probability distribution. The correspondence between the original belief model and the approximate finite model is governed by a nearest neighbor mapping under the bounded Lipschitz metric \( \rho_{BL} \).

In this section, we will construct a finite state space by quantizing the belief state space so that the approximate finite model is obtained using only a finite memory.

Our construction builds on but significantly differs from the approach in ([90, 93]). As we will explain, we cannot afford to use uniform quantization for this method, which was a crucial tool in ([90] 93).
As we discussed in the previous section, we can write the infinite horizon cost as

\[ J_\beta(T, Q, \gamma, \mu) = \sum_{t=0}^{\infty} \beta^t E_\mu \left[ \tilde{c}(\pi_t, \gamma(\pi_t)) \right] \]

\[ = \sum_{t=0}^{N-1} \beta^t E_\mu \left[ \tilde{c}(\pi_t, \gamma(\pi_t)) \right] + \sum_{t=N}^{\infty} \beta^t E_\mu \left[ \tilde{c}(\pi_t, \gamma(\pi_t)) \right]. \]

Now we focus on the second term:

\[ \sum_{t=N}^{\infty} \beta^t E_\mu \left[ \tilde{c}(\pi_t, \gamma(\pi_t)) \right] \]

\[ = \sum_{t=N}^{\infty} \beta^t E_\mu \left[ \tilde{c} \left( P^{\mu,\gamma}(X_t \in \cdot | Y_{[0,t]}, U_{[0,t-1]}), \gamma(P^{\mu,\gamma}(X_t \in \cdot | Y_{[0,t]}, U_{[0,t-1]})) \right) \right]. \]

Notice that for any time step \( t \geq N \) and for a fixed observation realization sequence \( y_{[0,t]} \) and for a fixed control action sequence \( u_{[0,t-1]} \), the state process can be viewed as

\[ P^{\mu}(X_t \in \cdot | Y_{[0,t]}, U_{[0,t-1]} = u_{[0,t-1]}) = P^{\pi_{t-N}}(X_t \in \cdot | Y_{[t-N,t]}, U_{[t-N,t-1]} = u_{[t-N,t-1]}) \]

where

\[ \pi_{t-N}(\cdot) = P^{\mu}(X_{t-N} \in \cdot | Y_{[0,t-N-1]} = y_{[0,t-N-1]}, U_{[0,t-N-1]} = u_{[0,t-N-1]}). \]

That is, we can view the state as the Bayesian update of \( \pi_{t-N} \), the predictor at time \( t - N \), using the observations \( Y_{t-N}, \ldots, Y_t \). Notice that with this representation only the most recent \( N \) observation realizations are used for the update and the past information of the observations is embedded in \( \pi_{t-N} \).
5.3. METHOD I: BELIEF-SPACE QUANTIZATION BASED ON
FINITE-MEMORY

Hence, we can view the state space for time stages $t \geq N$ as

$$\mathcal{Z} = \{ P^\pi(X_N \in \cdot | Y_{[0,N]}, U_{[0,N-1]}); \pi \in \mathcal{P}(X), Y_{[0,N]} \in Y^{N+1}, U_{[0,N-1]} \in U^N \}$$

Consider the following finite set defined by, for a fixed probability measure $\hat{\pi} \in \mathcal{P}(X)$,

$$\mathcal{Z}^N_{\hat{\pi}} := \{ P^{\hat{\pi}}(X_N \in \cdot | Y_{[0,N]}, U_{[0,N-1]}); Y_{[0,N]} \in Y^{N+1}, U_{[0,N-1]} \in U^N \} . \quad (5.3)$$

Define the map $F : \mathcal{Z} \rightarrow \mathcal{Z}^N_{\hat{\pi}}$, which will serve as the quantizer, with

$$F(z) := \arg \min_{y \in \mathcal{Z}^N_{\hat{\pi}}} \rho_{BL}(z, y) \quad (5.4)$$

This map separates the set $\mathcal{Z}$ into $|Y|^{N+1} \times |U|^N$ sets, accordingly this quantizes the set of probability measures.

To complete our approximate controlled Markov model, we now define the one-stage cost function $c^N : \mathcal{Z}^N_{\hat{\pi}} \times U \rightarrow [0, \infty)$ and the transition probability $\eta^N$ on $\mathcal{Z}^N_{\hat{\pi}}$ given realizations in $\mathcal{Z}^N_{\hat{\pi}} \times U$: For a given $z_i = P^{\hat{\pi}}(X_N \in \cdot | y_{[0,N]}^i, u_{[0,N-1]}^i)$ and control action $u$

$$c^N(z_i, u) = c^N(P^{\hat{\pi}}(X_N \in \cdot | y_{[0,N]}^i, u_{[0,N-1]}^i), u) := \tilde{c}(P^{\hat{\pi}}(X_N \in \cdot | y_{[0,N]}^i, u_{[0,N-1]}^i), u),$$

where $\tilde{c}$ is defined in (4.5) and

$$\eta^N(\cdot|z_i, u) = \eta_N(\cdot|P^{\hat{\pi}}(X_N \in \cdot | y_{[0,N]}^i, u_{[0,N-1]}^i), u)) := F \ast \eta(\cdot|P^{\hat{\pi}}(X_N \in \cdot | y_{[0,N]}^i, u_{[0,N-1]}^i), u),$$

where

$$F \ast \eta(z' | z, u) = \eta(\{ z \in \mathcal{P}(X) : F(z) = z' \}| z, u),$$
5.3. METHOD I: BELIEF-SPACE QUANTIZATION BASED ON FINITE-MEMORY

for all \( z' \in \mathcal{Z}^N \).

We thus have defined a finite state MDP with the state space \( \mathcal{Z}^N \), action space \( \mathcal{U} \), cost function \( c_N \) and the transition probability \( \eta_N \).

An optimal policy \( \gamma^*_N \) for this finite state model is a function taking values from the finite state space and hence at some time step \( t \geq N \) it only uses \( N \) most recent observations and control action variables that is, \( \gamma^*_N \) is a measurable function of the information set \( I^N_t \) defined in (5.1) for all \( t \geq N \).

**Remark 5.3.1.** As we have noted before, the complexity of POMDPs in general arises from the structure of the belief state space \( \mathcal{Z} \) which is a set of probability measures on \( X \). This set is always uncountable and needs to be associated with proper topologies to make the analysis feasible. Approximations for POMDPs are usually done by choosing a finite subset, say \( \hat{\mathcal{Z}} \), of the belief state space \( \mathcal{Z} \) (e.g. [99, 84, 93, 119, 120, 101, 118]), and finding an approximate MDP model for this finite set. To choose the finite set, the aforementioned works use a uniform quantization scheme, in various topologies on \( \mathcal{Z} \). In other words, the quantization is made such that for any \( z \in \mathcal{Z} \), there exists an element \( \hat{z} \in \hat{\mathcal{Z}} \) with \( \| z - \hat{z} \| \leq \epsilon \) for a fixed \( \epsilon > 0 \). The metric to measure distances of the belief states varies for different works, although for finite \( X \), \( L_1 \) distance of distributions is what is used in general for the quantization of \( \mathcal{Z} \), which coincides with total variation and weak convergence topology on \( \mathcal{Z} \) when \( X \) is finite; for general \( X \), a more appropriate and natural topology is the weak convergence topology on \( \mathcal{Z} \) which is what we use in this section since \( \rho_{BL} \) metrizes the weak convergence.

In this section, instead of quantizing \( \mathcal{Z} \) directly and uniformly, we use finite window information variables \( I^N_t \)'s to construct the finite subset of \( \mathcal{Z} \) since our goal in this section is to analyze the effect of the window size on the approximation performance. That is, we
5.3. METHOD I: BELIEF-SPACE QUANTIZATION BASED ON FINITE-MEMORY

use the finite set

\[ \mathcal{Z}_*^N := \left\{ P_\pi^* (X_N \in \cdot | Y_{[0,N]}, U_{[0,N-1]}); Y_{[0,N]} \in Y_{N+1}^N, U_{[0,N-1]} \in U^N \right\} \]

constructed using \( Y_{[0,N]}, U_{[0,N-1]} \). For this set, we cannot afford a uniform discretization scheme. A uniform quantization would mean that for a fixed \( \epsilon > 0 \)

\[ \rho_{BL} \left( P_\pi^* (X_N \in \cdot | y_{[0,N]}, u_{[0,N-1]}), P_\pi^* (X_N \in \cdot | y_{[0,N]}, u_{[0,N-1]}) \right) < \epsilon \]

uniformly for any \( \pi \in \mathcal{Z} \) and for any \( y_{[0,N]} \in Y_{N+1}^N, u_{[0,N-1]} \in U^N \). However, this is in general inapplicable for filter stability problems as it requires for the processes with different starting points uniformly for any realizations of information variables (even for highly unlikely ones). That is why, we follow a different approach and show that we do not have to force uniform quantization, and the error of approximate value functions can be related to the expected error of the form

\[ E^\gamma_* [\rho_{BL} \left( P_\pi (X_N \in \cdot | Y_{[0,N]}, U_{[0,N-1]}), P_\pi (X_N \in \cdot | Y_{[0,N]}, U_{[0,N-1]}) \right)] \]

which in turn can be bounded using Theorem 5.2.1. Our technical analysis, accordingly, is slightly more tedious at the benefit of arriving at a practical and intuitive finite-memory method whose near optimality is rigorously established.

5.3.1 Approximation Error Analysis and Rates of Convergence

An optimal policy for the constructed finite model, \( \gamma^*_N \), can be extended to \( \mathcal{P}(X) \) and can be used for the original MDP. The next result, which is related to the construction given
5.3. METHOD I: BELIEF-SPACE QUANTIZATION BASED ON
FINITE-MEMORY

in ([90, Thm 4.38]) (see also [93]), provides a mismatch error for using this policy. This result is going to be the key supporting tool for the main theorem of the section, which will be presented immediately after. The proof requires multiple technical lemmas and are presented in Section 5.7.1.

Assumption 5.3.1. • \( \rho_{BL}(\eta(z, u), \eta(z', u)) \leq \alpha_z \rho_{BL}(z, z') \) (see Theorem 4.4.3).

• \( |\tilde{c}(z, u) - \tilde{c}(z', u)| \leq \alpha_{\tilde{c}} \rho_{BL}(z, z') \) for some \( \alpha_{\tilde{c}} \) for all \( u \in U \).

Theorem 5.3.1. Under Assumption 5.3.1, for all \( z \in \mathcal{P}(\mathcal{X}) \) of the form

\[ z = \mathcal{P}^\pi(X_N \in \cdot | Y_{[0,N]}, u_{[0,N-1]}), \]

the following holds:

\[
\sup_{\gamma \in \Gamma} E \left[ J_\beta(\eta^*, \gamma_N, z) - J_\beta(\eta, \gamma, z) \bigg| Y_{[0,N]}, \gamma(Y_{[0,N-1]}) \right] \\
\leq K(\beta, \alpha_z, \alpha_{\tilde{c}}, \|\tilde{c}\|_{\infty}) \sup_{\pi \in \mathcal{P}^{\mathcal{X}}} \sup_{\gamma \in \Gamma} E^\gamma_{\pi} \left[ \rho_{BL} \left( \mathcal{P}^\pi(X_N \in \cdot | Y_{[0,N]}, U_{[0,N-1]}), \mathcal{P}^\hat{\pi}(X_N \in \cdot | Y_{[0,N]}, U_{[0,N-1]}) \right) \right].
\]

Where \( K(\beta, \alpha_z, \alpha_{\tilde{c}}, \|\tilde{c}\|_{\infty}) \) is a constant that depends on \( \beta, \alpha_z, \alpha_{\tilde{c}} \) and \( \|\tilde{c}\|_{\infty} \). (The exact expression for the constant can be found in the proof).

The proof of this theorem is rather long, and accordingly, it is presented in Section 5.7.1.

Before presenting the main result of the section, we provide further supporting results that will let us work with a probability density function.

Lemma 5.3.1. Assume that the transition kernel \( \mathcal{T}(dx_1|x_0, u_0) \) admits a density function \( f \) with respect to a reference measure \( \phi \) such that \( \mathcal{T}(dx_1|x_0, u_0) = f(x_1, x_0, u_0)\phi(dx_1) \). If
5.3. METHOD I: BELIEF-SPACE QUANTIZATION BASED ON FINITE-MEMORY

\[ |f(x_1, x_0, u_0) - f(x_1, x_0', u_0)| \leq \alpha_X |x_0 - x_0'| \text{ for all } x_1, x_0, x_0' \in X \text{ and } u_0 \in U \text{ then} \]

\[ \|T(\cdot|x_0, u_0) - T(\cdot|x_0', u_0)\|_{TV} \leq \alpha_X |x - x'|. \]

We note that using this result, assumptions of Theorem 4.4.3 can be expressed with the Lipschitz condition on the density function noted above. We now restate the assumptions that will be used for the main result.

**Assumption 5.3.2.**

- The transition kernel \( T(dx_1|x_0, u_0) \) admits a density function \( f \) with respect to a reference measure \( \phi \) such that \( T(dx_1|x_0, u_0) = f(x_1, x_0, u_0)\phi(dx_1) \).
- There exists some \( \alpha_X < \infty \) such that \( |f(x_1, x_0, u_0) - f(x_1, x_0', u_0)| \leq \alpha_X |x_0 - x_0'| \).
- There exists some \( \alpha_c < \infty \) such that for all \( u \in U \), \( |c(x, u) - c(x', u)| \leq \alpha_c |x - x'| \).
- \( \alpha := (1 - \tilde{\delta}(T))(2 - \delta(Q)) < 1 \).
- The transition kernel \( T \) is dominated, i.e. there exists a dominating measure \( \hat{\pi} \in \mathcal{P}(X) \) such that for every \( x \in X \) and \( u \in U \), \( T(\cdot|x, u) \ll \hat{\pi}(\cdot) \) that is \( T(\cdot|x, u) \) is absolutely continuous with respect to \( \hat{\pi} \) for every \( x \in X \) and \( u \in U \).

**Theorem 5.3.2.** Assume that we let the system start running for \( N \) time steps under a known policy \( \gamma \) (which may not be optimal), and the finite window policy starts acting after observing \( N \) information variables at time \( t = N \). Under Assumption 5.3.2, we have

\[ E_{\mu}^\gamma [J_\beta(\pi_N, T, \gamma_N^*) - J_\beta(\pi_{N-1}, T, \gamma_N^*)|Y_{[0,N]}, U_{[0,N-1]}] \leq K(\beta, \alpha_X, \alpha_c, \|c\|_\infty)\alpha^N. \]
where $\gamma^*_N$ is the optimal finite window policy and where $K(\beta, \alpha_X, \alpha_c, \|c\|_{\infty})$ is a constant that depends on $\beta, \alpha_X, \alpha_c$ and $\|c\|_{\infty}$ (The exact formula for the constant can be found in the appendix).

**Remark 5.3.2.** We note that Theorem 5.3.2 applies to all finite state, measurement, action models as long as $\alpha = (1 - \tilde{\delta}(T))(2 - \delta(Q)) < 1$. Example 5.2.1 shows how to calculate the Dobrushin coefficient for transition matrices in finite setups. All other conditions apply since all probability measures on a finite/countable set are majorized by a probability measure which places a positive mass on every single point. Notice that the third condition only requires that the difference in the cost is bounded for every fixed control action. Condition 1 and 5 coincide in the finite case.

**Remark 5.3.3.** We note that the error bound of the result is independent of the chosen $\hat{\pi}$. As we will see in the following proof, the first upper bound is a result of Theorem 5.3.1 which indeed depends on the $\hat{\pi}$ chosen by the user. However, thanks to Theorem 5.2.1 we can get a further upper bound on the error which is uniform over any $\hat{\pi}$ as long as $\hat{\pi}$ is a dominating measure.

**Proof of Theorem 5.3.2** When we reduce a partially observed MDP to a fully observed process, the initial state of the belief process becomes the Bayesian update of the prior distribution of the state process of the POMDP. Hence, we can write that

$$E_\mu^* \left[ J_\beta(\pi_{N-}, \gamma^*_N) - J_\beta(\pi_{N-}, \gamma^*) | Y_{[0,N]}, U_{[0,N-1]} \right]$$

$$= E_\mu^* \left[ J_\beta(\pi_N, \eta, \gamma^*_N) - J_\beta(\pi_N, \eta, \gamma^*) | Y_{[0,N]}, U_{[0,N-1]} \right].$$
Notice that, using Theorem 4.4.3, we have

$$\rho_{BL}(\eta(\cdot|z,u), \eta(\cdot|z',u)) \leq 5(1 + \alpha_X)\rho_{BL}(z,z')$$

and

$$|\tilde{c}(z,u) - \tilde{c}(z',u)| = \left| \int c(x,u)z(dx) - \int c(x,u)z'(dx) \right| \leq (\alpha_c + ||c||_\infty)\rho_{BL}(z,z').$$

Thus, we can use Theorems 4.4.3 and 5.3.1 to write

$$E^\gamma_\mu \left[ J_\beta(\pi_N, \eta, \gamma_N) - J_\beta(\pi_N, \eta, \gamma^*)|Y_{[0,N]}, U_{[0,N-1]} \right] \leq K(\beta, \alpha_X, \alpha_c, ||c||_\infty) \sup_{\pi \in P(X)} \sup_{\gamma \in \Gamma} E^\gamma_\pi \left[ \rho_{BL}(P_{\pi \gamma}(X \in \cdot|Y_{[0,N]}), P_{\hat{\pi} \gamma}(X \in \cdot|Y_{[0,N]})) \right].$$

We set $\hat{\pi}$ in Assumption 5.3.2 as our representative probability measure for the quantization of the belief space. Notice that by our choice of $\hat{\pi}$ is a dominating measure and therefore $\pi_t \ll \hat{\pi}$ for any time step $t$, where $\pi_t$ is the predictor at time $t$. Thus, Theorem 5.2.1 yields that under Assumption 5.3.2

$$E^\gamma_\pi \left[ ||P^\pi(\cdot|Y_{[0,N]}) - P^{\hat{\pi}}(\cdot|Y_{[0,N]})||_{TV} \right] \leq \alpha^N,$$

for any predictor $\pi$. We also have by definition that

$$\rho_{BL}(P^{\pi,\gamma}(\cdot|Y_{[0,N]}), P^{\hat{\pi},\gamma}(\cdot|Y_{[0,N]})) \leq ||P^{\pi,\gamma}(\cdot|Y_{[0,N]}) - P^{\hat{\pi},\gamma}(\cdot|Y_{[0,N]})||_{TV}.$$

Hence the result follows.
as our approximation technique uses information variables with size $N$, but for the initial $N$ steps, there is not enough observation or control action variables to make use of the approximate finite window policies. In the following result, for the first $N$ time steps, we use a control policy that is found with a policy iteration type argument where the terminal cost estimated with $\beta^N E[J_\beta(\gamma_N^*)]$ with $\gamma_N^*$ being the approximate finite window policy.

The policy is constructed as follows: Use the policy $\gamma_N^*$, after time $N$ unaltered, but modify the first $N$ time-stage policies, as a batch update, which can be solved via a finite dynamic programming algorithm.

$$
\tilde{\gamma}_0, \ldots, \tilde{\gamma}_{N-1} = \arg\min_{\gamma_1, \ldots, \gamma_{N-1}} E\left[\sum_{k=0}^{N-1} \beta^k c(x_k, u_k)\right] + \beta^N E\left[J_\beta(\gamma_N^*)\right| I_N
$$

where $I_N$ is the history by time $N$. We denote this policy by $\gamma^N := \{\tilde{\gamma}_0, \ldots, \tilde{\gamma}_{N-1}, \gamma_N^*, \gamma_N^*, \ldots\}$.

**Theorem 5.3.3.** Under Assumption 5.3.2 we have

$$
J_\beta(\mu, T, \gamma^N) - J_\beta^*(\mu, T) \leq K(\beta, \alpha_X, \alpha_c, \|c\|_\infty) \alpha^N \beta^N
$$

where $K(\beta, \alpha_X, \alpha_c, \|c\|_\infty)$ is a constant that depends on $\beta, \alpha_X, \alpha_c$ and $\|c\|_\infty$. (The exact formula for the constant can be found in the appendix).

**Proof.** Recall the optimal policy for the finite model constructed in Section 5.3 is denoted by $\gamma_N^*$. Notice that $\gamma_N^*$ is a stationary policy and optimal for any initial point since it solves the discounted cost optimality equation ([90]).

Now, we construct the following policy. Use the policy $\gamma_N^*$, after time $N$ unaltered, but modify the first $N$ time-stage policies, as a batch update, which can be solved via a finite dynamic programming algorithm.
\[ \tilde{\gamma}_0, \ldots, \tilde{\gamma}_{N-1} = \arg\min_{\gamma_1, \ldots, \gamma_{N-1}} E \left[ \sum_{k=0}^{N-1} \beta^k c(x_k, u_k) \right] + \beta^k E \left[ J_{\beta}(\gamma^*_N) | I_N \right] \]

where \(I_N\) is the history by time \(N\). We denote this policy by \(\gamma_N := \{\tilde{\gamma}_0, \ldots, \tilde{\gamma}_{N-1}, \gamma^*_N, \gamma^*_N, \ldots \}.\)

Note that we denote the true optimal policy for the original model by \(\gamma^*\). We now define the policy \(\gamma^+ := \{\gamma^*_{[0,N-1]}, \gamma^*_N, \gamma^*_N, \ldots \}:\) Apply \(\gamma^*\) until time \(N\), and then use our finite window policy \(\gamma^*_N\). Note that this policy is not practical as it assumes that the controller already knows the true optimal policy \(\gamma^*\), however, we will make use this hypothetical policy for the analysis.

From the way \(\gamma^N\) is constructed, we clearly, \(J_{\beta}(\gamma^N) \leq J_{\beta}(\gamma^+)\). And thus,

\[ J_{\beta}(\gamma^+) - J_{\beta}(\gamma^*) \leq J_{\beta}(\gamma^+) - J_{\beta}(\gamma^*) \]

Then, we start analyzing the error term: because \(\gamma^+\) and \(\gamma^*\) use the same policy before time \(N\), we have

\[ J_{\beta}(\mu, \mathcal{T}, \gamma^+) - J_{\beta}(\mu, \mathcal{T}, \gamma^*) \]

\[ = \sum_{t=N}^{\infty} \beta^t E_{\mu} \left[ c(X_t, \gamma^*_N(Y_{[t-N,N]}, U_{[t-N,t-1]})) \right] - \sum_{t=N}^{\infty} \beta^t E_{\mu} \left[ c(X_t, \gamma^*(Y_{[0,t]}, U_{[0,t-1]})) \right] \]

\[ = \beta N \sum_{t=N}^{\infty} \beta^{t-N} E_{\mu}^* \left[ E_{\mu} \left[ c(X_t, \gamma^*_N(Y_{[t-N,N]}, U_{[t-N,t-1]})) \right] \right. \]

\[ - E_{\mu} \left[ c(X_t, \gamma^*(Y_{[0,t]}, U_{[0,t-1]})) \right] | Y_{[0,N]}, U_{[0,N-1]} ] \]

\[ = \beta N E_{\mu}^* \left[ J_{\beta}(\pi_{N_-}, \mathcal{T}, \gamma^*_N) - J_{\beta}(\pi_{N_-}, \mathcal{T}, \gamma^*) | Y_{[0,N]}, U_{[0,N-1]} ] \right] \quad (5.5) \]

The last step follows from the observation that conditioning on the observations and control actions \(Y_{[0,N]}, U_{[0,N-1]}\), the state process can be thought as if it starts at time \(t = N\) whose
prior measure is $\pi_N$ and from the fact that the probability measure of $Y_{[0,N]}, U_{[0,N-1]}$ is determined by the initial measure $\mu$ and the policy $\gamma^*$ since $\gamma_t^N = \gamma_t^*$ for $t \leq N$.

The result then follows from Theorem 5.3.2.

Remark 5.3.4. The above results suggest that the loss occurring from applying a finite window policy is mainly controlled by the term

$$\sup_{\pi} \sup_{\gamma} E_{\pi}^{\gamma} \left[ BL \left( \mu_{[0,N]}, \mu_{[0,N]} \right) \right]$$

that is how fast two different belief-state processes forget their initial priors when fed with the same observations/control actions. Thus, any bound for this term directly applies to the main bound for the loss caused by the finite window policy. This term is related to the filter stability problem and our approximation results point out the close relation between filter stability and the performance of finite window policies. In a way, the main result of this section can be seen as relating the finite window approximations for POMDPs to the filter stability problem.

To bound the term (5.6), we use Theorem 5.2.1 to achieve a geometric convergence rate in the window size for a controlled setup. However, we should note that, Theorem 5.2.1 provides only a sufficient condition to bound the filter stability term geometrically fast in the total variation distance. This result can be seen as too strong, if one is only interested in making this $BL$ distance smaller with increasing window size. In fact, as we will see in simulation examples, even when the assumptions of Theorem 5.2.1 are not satisfied, the filter stability term still converges to 0. In the literature, there are various set of assumptions to achieve filter stability. Two main approaches have been:
5.3. METHOD I: BELIEF-SPACE QUANTIZATION BASED ON
FINITE-MEMORY

- The transition kernel is in some sense sufficiently ergodic, forgetting the initial measure and therefore passing this insensitivity (to incorrect initializations) on to the filter process.

- The measurement channel provides sufficient information about the underlying state, allowing the filter to track the true state process. This approach is typically based on martingale methods and accordingly does not often lead to rates of convergence for the filter stability problem, but only asymptotic filter stability.

The result we use in this chapter (Theorem 5.2.1) provides geometric filter stability, using a joint contraction property of the Bayesian filter update and measurement update steps through the Dobrushin coefficient. However, when these requirements are not satisfied, the filter stability can be checked via different assumptions from the literature. For a comprehensive review on filter stability in the control-free case, see (127). In the controlled case, recent studies include (77, 78).

The following example studies an additive Gaussian model (not necessarily linear) and provides different parameters for the condition $(1 - \delta(T)) \times (2 - \delta(\bar{Q})) < 1$ to hold.

**Example 5.3.1.** Consider a system where $X = Y = \mathbb{R}$ and the transition and measurement kernels are given by

$$x_{n+1} = f(x_n, u_n) + N(0, \sigma_t^2), \quad y_n = g(x_n) + N(0, \sigma_q^2)$$

where the functions $f$ and $g$ are measurable and bounded such that $f(x, u) \in [-t, t]$ and $g(x) \in [-q, q]$.

Note that, in our results we assume and present our results from finite $Y$. Therefore, to make the example compatible with our results, we discretize the observation space $Y$. 
We provide two discretization schemes one with $\hat{Y}_1 = \{-q, q\}$ and the other with $\hat{Y}_2 = \{-q, 0, q\}$. For discretization, we simply use a nearest neighbor mapping.

First, we study the observation space $\hat{Y}_1 = \{-q, q\}$. Using the nearest neighbor mapping, we simply have that

$$\hat{y}_n = -q \text{ if } y_n = g(x_n) + N(0, \sigma_q^2) \leq 0,$$
$$\hat{y}_n = q \text{ if } y_n = g(x_n) + N(0, \sigma_q^2) > 0$$

We then can write the following for the transition kernel $T(\cdot| x, u) \in P(X)$

$$T(dx_{n+1}|x_n, u_n) \sim N(f(x_n, u_n), \sigma_t^2)$$

and for the compound channel $\hat{Q}(\cdot|x) \in P(\hat{Y}_1)$:

$$\hat{Q}(q|x_n) = Pr(N(g(x_n), \sigma_q^2) > 0), \quad \hat{Q}(-q|x_n) = Pr(N(g(x_n), \sigma_q^2) \leq 0).$$

For these kernels, the Dobrushin coefficients can be calculated as

$$\delta(T) = 2Pr(N(t, \sigma_t^2) \leq 0), \quad \delta(\hat{Q}) = 2Pr(N(q, \sigma_q^2) \leq 0).$$

Notice that these probabilities are fully determined by the ratio of the mean and standard deviation of the Gaussian in question, $\sigma_t/t$ and $\sigma_q/q$. The higher the ratio, the higher the Dobrushin coefficient. Below, we see a list of the ratio of the transition kernel and lowest possible ratio of the measurement kernel such that $(1 - \delta(T)) \times (2 - \delta(\hat{Q})) < 1$. If the ratio of $\sigma_q/q$ is higher than the stated value, we will get exponential stability for the given transition kernel. Note that if $\sigma_t/t > 1.5$, then $\delta(T) > 0.5$ which makes
(1 − δ(T)) × (2 − δ(ˆQ)) < 1 regardless of the channel for which we use ‘any’ in table to indicate that any channel would lead to exponential filter stability.

<table>
<thead>
<tr>
<th>σ_q^2</th>
<th>1.5</th>
<th>1.4</th>
<th>1.3</th>
<th>1.2</th>
<th>1.1</th>
<th>1.0</th>
<th>0.9</th>
<th>0.8</th>
<th>0.7</th>
<th>0.6</th>
<th>0.5</th>
<th>0.4</th>
<th>0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>\frac{\sigma_q^2}{q}</td>
<td>any</td>
<td>0.6</td>
<td>0.8</td>
<td>1.01</td>
<td>1.3</td>
<td>1.65</td>
<td>2.13</td>
<td>3.25</td>
<td>5.5</td>
<td>8.0</td>
<td>20.0</td>
<td>70.0</td>
<td>1000.0</td>
</tr>
<tr>
<td>δ(T)</td>
<td>0.50</td>
<td>0.48</td>
<td>0.44</td>
<td>0.40</td>
<td>0.36</td>
<td>0.32</td>
<td>0.27</td>
<td>0.21</td>
<td>0.15</td>
<td>0.10</td>
<td>0.05</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>δ(ˆQ)</td>
<td>any</td>
<td>0.1</td>
<td>0.21</td>
<td>0.32</td>
<td>0.44</td>
<td>0.54</td>
<td>0.64</td>
<td>0.76</td>
<td>0.86</td>
<td>0.90</td>
<td>0.96</td>
<td>0.99</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 5.1: Approximate minimum ratio of $\frac{\sigma_q^2}{q}$ for exponential filter stability for $\hat{Y}_1 = \{-q, q\}$

We now analyze the problem for the observation space $\hat{Y}_2 = \{-q, 0, q\}$. For this set, the nearest neighbor mapping yields that

\[
\hat{y}_n = \begin{cases} 
-q & \text{if } y_n = g(x_n) + N(0, \sigma_q^2) \leq -\frac{q}{2}, \\
q & \text{if } y_n = g(x_n) + N(0, \sigma_q^2) > \frac{q}{2}, \\
0 & \text{else.}
\end{cases}
\]

For the compound channel, we then have

\[
\hat{Q}(q|x_n) = Pr(N(g(x_n), \sigma_q^2) > q) = Pr(N(g(x_n), \sigma_q^2) \leq -q), \\
\hat{Q}(-q|x_n) = Pr(N(g(x_n), \sigma_q^2) > q) + Pr\left(\frac{q}{2} \leq N(g(x_n), \sigma_q^2) \geq -q\right), \\
\hat{Q}(0|x_n) = Pr\left(\frac{q}{2} \leq N(g(x_n), \sigma_q^2) \geq -q\right).
\]

For these kernels, the Dobrushin coefficients can be calculated as

\[
\delta(T) = 2Pr(N(t, \sigma_t^2) \leq 0),
\]
\[
\delta(\hat{Q}) = 2Pr(N(q, \sigma_q^2) \leq -\frac{q}{2}) + Pr\left(\frac{-q}{2} < N(q, \sigma_q^2) < \frac{q}{2}\right).
\]
Below, we now see a list of the ratio of the transition kernel and lowest possible ratio of the measurement kernel such that \((1 - \delta(\mathcal{T})) \times (2 - \delta(\mathcal{Q})) < 1\) for the observation space \(\mathcal{Y}_2 = \{-q, 0, q\}\).

<table>
<thead>
<tr>
<th>(\frac{\sigma_q}{T})</th>
<th>1.5</th>
<th>1.4</th>
<th>1.3</th>
<th>1.2</th>
<th>1.1</th>
<th>1.0</th>
<th>0.9</th>
<th>0.8</th>
<th>0.7</th>
<th>0.6</th>
<th>0.5</th>
<th>0.4</th>
<th>0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\frac{\delta_q}{q})</td>
<td>any</td>
<td>0.39</td>
<td>0.6</td>
<td>0.85</td>
<td>1.2</td>
<td>1.54</td>
<td>2.1</td>
<td>3.2</td>
<td>5.9</td>
<td>8.0</td>
<td>20.0</td>
<td>80.0</td>
<td>1000.0</td>
</tr>
<tr>
<td>(\delta(\mathcal{T}))</td>
<td>0.50</td>
<td>0.48</td>
<td>0.44</td>
<td>0.40</td>
<td>0.36</td>
<td>0.32</td>
<td>0.27</td>
<td>0.21</td>
<td>0.15</td>
<td>0.10</td>
<td>0.05</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>(\delta(\mathcal{Q}))</td>
<td>any</td>
<td>0.1</td>
<td>0.21</td>
<td>0.32</td>
<td>0.44</td>
<td>0.54</td>
<td>0.64</td>
<td>0.76</td>
<td>0.86</td>
<td>0.90</td>
<td>0.96</td>
<td>0.99</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 5.2: Approximate minimum ratio of \(\frac{\sigma_q}{q}\) for exponential filter stability for \(\mathcal{Y}_2 = \{-q, 0, q\}\)

5.4 Method II: Finite Window Memory as Approximate POMDP State

We now present our second approximation method for the construction of the finite belief MDPs. This method also makes use of a finite belief state space that is constructed with reachable posterior distributions under finite history and a fixed initial distribution. However, different from the first method used in Section 5.3, here the correspondence between the finite model and the original model is governed by matching the belief variables with same recent history. This procedure will be made more rigorous in the following parts.

Compared to the method used in Section 5.3, the method used in this section will naturally result in coarser error bounds because the mapping to the finite state space is done via matching the recent history instead of using a nearest neighbor map. However, this method will prove to have practical and computational advantages as we will also see in Chapter 6.

5.4.1 An Alternative Finite Window Belief-MDP Reduction

In this section we construct an alternative fully observed MDP reduction (alternative to Section 4.1) with the condition that the controller has observed at least \(N\) information
variables, using the predictor from $N$ stages earlier and the most recent $N$ information variables (that is, measurements and actions). This new construction allows us to highlight the most recent information variables and *compress* the information coming from the past history via the predictor as a probability measure valued variable. In what follows, we will sometimes consider the case with $N = 1$ for some of the proofs to make the presentation less complicated. The general case follows from identical arguments.

For the remainder of the section, to emphasize the prior distribution of the starting state variable, we will use the following notation for conditional probabilities on state and observation variables.

**Definition 5.4.1.** Assume that the initial state $X_0$ has a prior distribution $\mu \in \mathcal{P}(X)$. Then, for the conditional distribution of $X_t$ given the past observation and action variables $\{y_t, \ldots, y_0\}, \{u_{t-1}, \ldots, u_0\}$ we define

$$P^\mu(X_t \in \cdot | y_t, \ldots, y_0, u_{t-1}, \ldots, u_0) := Pr(X_t \in \cdot | y_t, \ldots, y_0, u_{t-1}, \ldots, u_0).$$

Given that $x_0$ has a prior distribution $\mu \in \mathcal{P}(X)$, we define the following for the conditional distribution of $Y_t$ given the past observation and action variables $\{y_{t-1}, \ldots, y_0\}, \{u_{t-1}, \ldots, u_0\}$

$$P^\mu(Y_t \in \cdot | y_{t-1}, \ldots, y_0, u_{t-1}, \ldots, u_0) := Pr(Y_t \in \cdot | y_{t-1}, \ldots, y_0, u_{t-1}, \ldots, u_0).$$

Consider the following state variable at time $t$:

$$z_t = (\pi_{t-N}, I_t^N)$$

(5.7)
where, for $N \geq 1$

$$
\pi_{t-N} = Pr(X_{t-N} \in \cdot | y_{t-N-1}, \ldots, y_0, u_{t-N-1}, \ldots, u_0),
$$

$$
I_t^N = \{y_t, \ldots, y_{t-N}, u_{t-1}, \ldots, u_{t-N}\}
$$

and $I_t^N = y_t$ for $N = 0$ with $\mu$ being the prior probability measure on $X_0$. The state space with this representation is $\hat{Z} = \mathcal{P}(X) \times Y^{N+1} \times U^N$ where we equip $\hat{Z}$ with the product topology where we consider the weak convergence topology on the $\mathcal{P}(X)$ coordinate and the usual (coordinate) topologies on $Y^{N+1} \times U^N$ coordinates.

This new state representation can be mapped to the belief state $z_t$, which is defined as

$$
z_t := Pr\{X_t \in \cdot | Y_0, \ldots, Y_t, U_0, \ldots, U_{t-1}\} \in \mathcal{P}(X).
$$

Consider the map $\psi : \hat{Z} \rightarrow \mathcal{P}(X)$, for some $\hat{z}_t = (\pi_{t-N}^-, I_t^N)$

$$
\psi(\hat{z}_t) = \psi(\pi_{t-N}^-, I_t^N) = P^{\pi_{t-N}^-}(X_t \in \cdot | I_t^N) = P^{\pi_{t-N}^-}(X_t \in \cdot | y_t, \ldots, y_{t-N}, u_{t-N-1}, \ldots, u_{t-N-1}) = Pr(X_t \in \cdot | y_t, \ldots, y_0, u_{t-1}, \ldots, u_0) = z_t
$$

such that the map $\psi$ acts as a Bayesian update of $\pi_{t-N}^-$ using $I_t^N$. Using this map, we can define the stage-wise cost function and the transition probabilities. Consider the new cost function $\hat{c} : \hat{Z} \times U \rightarrow \mathbb{R}$, using the cost function $\tilde{c}$ of the belief MDP (defined in (4.5)) such that

$$
\hat{c}(\hat{z}_t, u_t) = \hat{c}(\pi_{t-N}^-, I_t^N, u_t) = \tilde{c}(\psi(\pi_{t-N}^-, I_t^N), u_t)
$$

$$
= \int_X c(x_t, u_t) P^{\pi_{t-N}^-}(dx_t | y_t, \ldots, y_{t-N}, u_{t-1}, \ldots, u_{t-N}). \tag{5.8}
$$
Furthermore, we can define the transition probabilities as follows: for some $A \in \mathcal{B}(\hat{Z})$ such that

$$A = B \times \{\hat{y}_{t-N+1}, \hat{u}_t, \ldots, \hat{u}_{t-N+1}\}, \quad B \in \mathcal{B}(\mathcal{P}(X))$$

we write

$$Pr(\hat{z}_{t+1} \in A | \hat{z}_t, \ldots, \hat{z}_0, u_t, \ldots, u_0)$$

be defined as

$$G(\pi_{t-N}^-; y_{t-N}, u_{t-N}) = G(P^\mu(X_{t-N} \in \cdot | y_{t-N-1}, \ldots, y_0, u_{t-N-1}, \ldots, u_0), y_{t-N}, u_{t-N})$$

$$= P^\mu(X_{t-N+1} \in \cdot | y_{t-N}, \ldots, y_0, u_{t-N}, \ldots, u_0).$$

Hence, $\hat{\eta}$ defines a controlled transition model for the new states $\hat{z}_{t+1} \in \hat{Z}$. Then, we have a proper fully observed MDP, with the cost function $\hat{c}$, transition kernel $\hat{\eta}$ and the state space $\hat{Z}$.

Note that any policy $\phi: \mathcal{P}(X) \rightarrow U$ defined for the belief MDP, can be extended to the newly defined finite window belief-MDP using the map $\psi$, and defining $\hat{\phi} := \phi \circ \psi$ such
5.4. METHOD II: FINITE WINDOW MEMORY AS APPROXIMATE POMDP

that

\[ \hat{\phi}(\hat{z}) = \phi(\psi(z)) . \]

Thus, if an optimal policy can be found for the belief MDP, say \( \phi^* \), the policy \( \hat{\phi}^* = \phi^* \circ \psi \) is an optimal policy for the newly defined MDP.

We now write the discounted cost optimality equation for the newly constructed finite window belief MDP. Note that with the alternative approach the state \( \hat{z} \) can only be written, if we have at least \( N \) information variables. Therefore, given that the decision maker observed at least \( N \) information variables, we write the following fixed point equation

\[
J^*_\beta(\hat{z}) = \min_{u \in U} \left( \hat{c}(\hat{z}, u) + \beta \int J^*_\beta(\hat{z}_1) \hat{\eta}(d\hat{z}_1 | \hat{z}, u) \right) .
\]

We can rewrite this fixed point equation in a different form, for notational ease assume \( N = 1 \). If \( \hat{z} \) has the form \((\pi_0^{-}, y_1, y_0, u_0)\), then we can rewrite

\[
J^*_\beta(\pi_0^{-}, y_1, y_0, u_0) = \min_{u_1 \in U} \left( \hat{c}(\pi_0^{-}, y_1, y_0, u_0, u_1) + \beta \sum_{y_2 \in Y} J^*_\beta(\pi_1^{-}(\pi_0^{-}, y_0, u_0), y_2, y_1, u_1) P^{\pi_0^{-}}(y_2|y_1, y_0, u_1, u_0) \right) .
\]

(5.9)

This representation will play an important role in the analysis of the problem. Note that the policy \( \hat{\phi}^* = \phi^* \circ \psi \) satisfies this fixed point equation.

The following fixed point equation can also be defined for any policy \( \hat{\phi} : \hat{Z} \to U \)

\[
J_\beta(\hat{z}, \hat{\phi}) = \hat{c}(\hat{z}, \hat{\phi}(\hat{z})) + \beta \int J_\beta(\hat{z}_1, \hat{\phi}) \hat{\eta}(d\hat{z}_1 | \hat{z}, \hat{\phi}(\hat{z}))
\]
where \( J_\beta(\hat{z}, \hat{\phi}) \) denotes the value function under the policy \( \hat{\phi} \) for the initial point \( \hat{z} \).

### 5.4.2 Approximation of the Finite Window Belief-MDP

We now approximate the MDP constructed in the previous section. Consider the following set \( \hat{Z}_N^\pi \) for a fixed \( \hat{\pi} \in \mathcal{P}(X) \) (recall (5.3)),

\[
\hat{Z}_N^\pi = \left\{ (\hat{\pi}, y_{[0,N]}, u_{[0,N-1]}): y_{[0,N]} \in Y_{N+1}^N, u_{[0,N-1]} \in U^N \right\} \tag{5.10}
\]

such that the state at time \( t \) is \( \hat{z}_t^N = (\hat{\pi}, I_t^N) \). Compared to the state \( \hat{z}_t = (\pi_{t-N}, I_t^N) \) defined in (5.7), this approximate model uses \( \hat{\pi} \) as the predictor, no matter what the real predictor at time \( t - N \) is.

The cost function is defined in usual manner so that

\[
\hat{c}(\hat{z}_t^N, u_t) = \hat{c}(\hat{\pi}, I_t^N, u_t) = \hat{c}(\phi(\hat{\pi}, I_t^N), u_t)
= \int_X c(x_t, u_t) P^\hat{\pi}(dx_t | y_t, \ldots, y_{t-N}, u_{t-1}, \ldots, u_{t-N}).
\]

We define the controlled transition model by

\[
\hat{\eta}^N(\hat{z}_{t+1}^N | \hat{z}_t^N, u_t) = \hat{\eta}^N(\hat{\pi}, I_{t+1}^N | \hat{\pi}, I_t^N, u_t) := \hat{\eta}\left( \mathcal{P}(X), I_{t+1}^N | \hat{\pi}, I_t^N, u_t \right). \tag{5.11}
\]

For simplicity, if we assume \( N = 1 \), then the transitions can be rewritten for some \( I_{t+1}^N = (\hat{y}_{t+1}, \hat{\pi}, \hat{u}_t, \hat{u}_t) \) and \( I_t^N = (y_t, y_{t-1}, u_{t-1}) \)

\[
\hat{\eta}^N(\hat{\pi}, \hat{y}_{t+1}, \hat{\pi}, \hat{u}_t | \hat{\pi}, y_t, y_{t-1}, u_{t-1}, u_t) = \hat{\eta}(\mathcal{P}(X), \hat{y}_{t+1}, \hat{\pi}, \hat{u}_t | \hat{\pi}, y_t, y_{t-1}, u_{t-1}, u_t)
= \mathbf{1}_{\{y_t = \hat{y}_t, u_t = \hat{u}_t\}} P^\hat{\pi}(\hat{y}_{t+1} | y_t, y_{t-1}, u_t, u_{t-1}). \tag{5.12}
\]
5.4. METHOD II: FINITE WINDOW MEMORY AS APPROXIMATE POMDP

Denoting the optimal value function for the approximate model by $J^N_\beta$, we can write the following fixed point equation

$$J^N_\beta(\hat{z}^N) = \min_{u \in U} \left( \hat{c}(\hat{z}^N, u) + \beta \sum_{\hat{z}_1^N \in \hat{Z}_k^N} J^N_\beta(\hat{z}_1^N) \hat{\eta}^N(\hat{z}_1^N | \hat{z}^N, u) \right). \quad (5.13)$$

By assuming $N = 1$ again, we can rewrite the fixed point equation for some $\hat{z}_0^N = (\hat{\pi}, y_1, y_0, u_0)$ as

$$J^1_\beta(\hat{\pi}, y_1, y_0, u_0) = \min_{u_1 \in U} \left( \hat{c}(\hat{\pi}, y_1, y_0, u_0, u_1) + \beta \sum_{y_2 \in Y} J^N_\beta(\hat{\pi}, y_2, y_1, u_1) P^{\hat{\pi}}(y_2 | y_1, y_0, u_1, u_0) \right). \quad (5.14)$$

Since everything is finite in this setup, we can assume the existence of an optimal policy $\phi^N$ that satisfies this fixed point equation. Note that both $J^N_\beta$ and $\phi^N$ are defined on the finite set $\hat{Z}_k^N$. However, we can simply extend them to the set $\hat{Z}$ by defining

$$\tilde{J}^N_\beta(\hat{z}) = J^N_\beta(\pi, y_1, y_0, u_0) := J^N_\beta(\hat{\pi}, y_1, y_0, u_0)$$

$$\tilde{\phi}^N(\hat{z}) = \phi^N(\pi, y_1, y_0, u_0) := \phi^N(\hat{\pi}, y_1, y_0, u_0)$$

for any $\hat{z} = (\pi, y_1, y_0, u_0) \in \hat{Z}$.

Another point to note is that the policy $\phi^N$ only uses most recent $N$ information variables to choose the control actions.

In what follows, we investigate the following approximation error terms

$$|\tilde{J}^N_\beta(\hat{z}) - J^*_\beta(\hat{z})|,$$

$$J_\beta(\hat{z}, \tilde{\phi}^N) - J^*_\beta(\hat{z}).$$
The first one is the difference between the optimal value function of the original model and that for the approximate model. The second term is the performance loss due to the policy calculated for the approximate model being applied to the true model.

### 5.4.3 Approximation Error Analysis and Rates of Convergence

#### Difference in the Value Functions

In this section, we study the difference $|\tilde{J}^N_\beta (\hat{z}) - J^*_\beta (\hat{z})|$.

Before the result, we introduce some notation. We first define the measurable policies with respect to the new state space $\hat{Z}$ by $\hat{\Gamma}$. That is, a policy $\hat{\gamma} \in \hat{\Gamma}$ is a sequence of control functions $\{\hat{\gamma}_t, t \in \mathbb{Z}_+\}$ such that $\hat{\gamma}_t$ is measurable with respect to the $\sigma$-algebra generated by the information variables $\{\hat{z}_0, \ldots, \hat{z}_t\}$.

We now define the following bounding term

$$L_t := \sup_{\hat{\gamma} \in \hat{\Gamma}} E^{\hat{\gamma}}_{\pi^0} \left[ \|P^{\pi^0} (X_{t+N} \in \cdot | Y_{[t,t+N]}, U_{[t,t+N-1]}) - P^\hat{\gamma} (X_{t+N} \in \cdot | Y_{[t,t+N]}, U_{[t,t+N-1]}) \|_{TV} \right]$$

(5.15)

which is the expected bound on the total variation distance between the posterior distributions of $X_{t+N}$ conditioned on the same observation and control action variables $Y_{[t,t+N]}, U_{[t,t+N-1]}$ when the prior distributions of $X_t$ are given by $\pi_t$ and $\hat{\pi}$. The expectation is with respect to the random realizations of $\pi_t$ and $Y_{[t,t+N]}, U_{[t,t+N-1]}$ under the true dynamics of the system when the prior distribution of $x_0$ is given by $\pi_0^\ominus$. This constant represents the bound on the distance of two processes with different starting points when they are updated with identical observation and action processes and under the same policy.

**Theorem 5.4.1.** For $\hat{z}_0 = (\pi_0^\ominus, I_0^N)$, if a policy $\hat{\gamma}$ acts on the first $N$ step of the process


which produces $I_0^N$, we then have

$$E^\gamma_{\pi_0} \left[ \left| J^N_{\beta}(\hat{z}_0) - J^*_{\beta}(\hat{z}_0) \right| | I_0^N \right] \leq \frac{\|c\|_\infty}{(1 - \beta)} \sum_{t=0}^{\infty} \beta^t L_t$$

where $L_t$ is defined as in (5.15).

**Proof.** The proof can be found in Section 5.7.6.

---

**Performance Loss due to Approximate Policy Being Applied to the True System**

We now study the difference $J_{\beta}(\hat{z}, \tilde{\phi}^N) - J^*_{\beta}(\hat{z})$ where $\tilde{\phi}^N$ is the optimal policy for the approximate model extended to the full space $\hat{Z}$.

**Theorem 5.4.2.** For $\hat{z}_0 = (\pi_0^N, I_0^N)$, with a policy $\gamma$ acting on the first $N$ steps

$$E^\gamma_{\pi_0} \left[ \left| J_{\beta}(\hat{z}_0, \tilde{\phi}^N) - J^*_{\beta}(\hat{z}_0) \right| | I_0^N \right] \leq \frac{2\|c\|_\infty}{(1 - \beta)} \sum_{t=0}^{\infty} \beta^t L_t.$$  

**Proof.** The proof can be found in Section 5.7.7.

---

**Exponential Convergence Rates under Filter Stability**

We now show that the term $L_t$ can be bounded using the following filter stability result. Recall that Theorem 5.2.1 states that under some sufficient conditions

$$E^{\mu, \gamma} \left[ \|\pi^\mu_n - \pi^\nu_n\|_{TV} \right] \leq 2\alpha^n.$$  

which holds uniformly for all $\mu \ll \nu$ where $\alpha := (1 - \tilde{\delta}(T))(2 - \delta(Q))$. Since $\tilde{\delta}(T)$ is a uniform Dobrushin coefficient over all control actions, the above bound is valid under any
control action process. Thus, under the conditions of Theorem 5.2.1, we have that

$$L_t = \sup_{\hat{\gamma} \in \hat{\Gamma}} E_{\hat{\gamma}}^{\hat{\pi}_0} \left[ \| P^{\pi_{t}}(X_{t+N} \in \cdot | Y_{[t,t+N]}, U_{[t,t+N-1]}) - P^{\hat{\pi}}(X_{t+N} \in \cdot | Y_{[t,t+N]}, U_{[t,t+N-1]}) \|_{T^V} \right]$$

$$\leq 2\alpha^N$$

for all $t$.

We can then state the following result:

**Theorem 5.4.3.** Assume the following holds:

- The transition kernel $T(dx_1|x_0, u_0)$ admits a density function $f$ with respect to a reference measure $\phi$ such that $T(dx_1|x_0, u_0) = f(x_1, x_0, u_0)\phi(dx_1)$.
- $\alpha := (1 - \bar{\delta}(T))(2 - \delta(Q)) < 1$.
- The transition kernel $T$ is dominated, i.e. there exists a dominating measure $\hat{\pi} \in \mathcal{P}(X)$ such that for every $x \in X$ and $u \in U$, $T(\cdot|x, u) \ll \hat{\pi}(\cdot)$ that is $T(\cdot|x, u)$ is absolutely continuous with respect to $\hat{\pi}$ for every $x \in X$ and $u \in U$.

Then, by choosing the dominating measure $\hat{\pi}$ for the construction of the approximate model, we have

$$E_{\hat{\gamma}}^{\hat{\pi}_0} \left[ \| \hat{J}_\beta^N(\hat{z}_0) - J_\beta^*(\hat{z}_0) \|_{I_0^N} \right] \leq \frac{2\|c\|_\infty}{(1 - \beta)^2} \alpha^N$$

$$E_{\hat{\gamma}}^{\hat{\pi}_0} \left[ \| J_\beta(\hat{z}_0, \hat{\phi}^N) - J_\beta^*(\hat{z}_0) \|_{I_0^N} \right] \leq \frac{4\|c\|_\infty}{(1 - \beta)^2} \alpha^N.$$ 

**Proof.** First note that using the dominating measure assumption, one can see that $\pi_t^{-} \ll \hat{\pi}$ for all $t$, where $\pi_t^{-}$ is the predictor at time $t$. 
Hence, with (5.16), we can write

\[ L_t \leq 2\alpha^N \]

for all \( t \), which concludes the proof using Theorem (5.4.2) and Theorem 5.4.1.

### 5.5 Comparison of Approximation Methods

In this chapter, we have presented two different approximation methods for POMDPs in Section 5.3 and Section 5.4 using finite window information variables. For both methods, the finite subset of the belief space is chosen in the same spirit. Namely, for Section 5.3, we have used

\[ \hat{Z}_N^\pi := \left\{ P^\pi(X_N \in \cdot | Y_{[0,N]}, U_{[0,N-1]}); Y_{[0,N]} \in Y_{N+1}^N, U_{[0,N-1]} \in U_{N}^N \right\} \]

and for Section 5.4, we have used

\[ \hat{Z}_N^\pi = \left\{ (\hat{\pi}, y_{[0,N]}, u_{[0,N-1]}); y_{[0,N]} \in Y_{N+1}^N, u_{[0,N-1]} \in U_{N}^N \right\} \]

In Section 5.3, note that the set is constructed using the posterior probability distributions of the form \( P^\pi(X_N \in \cdot | Y_{[0,N]}, U_{[0,N-1]}). \) We have equipped this set with the weak convergence topology as each of these elements is a probability measure. We then map the elements in the belief state space \( \hat{Z} \) to the finite subset using a nearest neighbor map choosing the closest element from \( \hat{Z}_N^\pi \) under \( \rho_{BL} \) metric which metrizes the weak convergence topology. The advantage of this method is that because of the nearest neighborhood map, one naturally arrives at a smaller approximation error. Furthermore, because of the
continuity properties of the nearest neighbor map under the $\rho_{BL}$ metric, one is able to work with the weak convergence topology, as such, we get an upper bound in terms of the $\rho_{BL}$ metric under the expectation, such that the bounding term is

$$E^\gamma_{\pi} \left[ \rho_{BL} \left( P^{\pi} (X_N \in \cdot | Y_{[0,N]}), P^\hat{\pi} (X_N \in \cdot | Y_{[0,N]}) \right) \right]$$

where the expectation is over the observation and action variables $Y_{[0,N]}, U_{[0,N-1]}$ which is a better bound than

$$E^\gamma_{\pi} \left[ \| P^{\pi} (X_N \in \cdot | Y_{[0,N]}) - P^\hat{\pi} (X_N \in \cdot | Y_{[0,N]}) \|_{TV} \right] \quad (5.17)$$

the bound we derived for the method used in Section 5.4 since the total variation distance always upper bounds the $\rho_{BL}$ metric.

Although, we get a better error bound in Section 5.3, since we directly work with the weak convergence topology on the set $Z^N_{\hat{\pi}}$, we can only differentiate between the probability measures in the set $Z^N_{\hat{\pi}}$ and we might lose the distinctions between different finite window information variables $Y_{[0,N]}, U_{[0,N-1]}$. In other words, different observation and control action realizations might lead to the same probability measure in the weak convergence topology. This issue might cause practical problems, because, the designer is always required to keep track of the original belief state to map it to the closest element $Z^N_{\hat{\pi}}$, instead of only keeping track of the information variables $Y_{[0,N]}, U_{[0,N-1]}$.

We have developed another approximation method in Section 5.4, which leads to a coarser error bound in (5.17) but results in a more practical way of approximation. Because of the issue explained in the last paragraph about losing the distinction between the most recent information variables, instead of putting a topology on the posterior distributions
直接，我们视元素在集合 $\mathring{Z}_N^{\pi}$ 作为预估器 $N$ 步后的组合和 $N$ 窗口信息变量，即我们视元素为 $(\hat{\pi}, I_t^N)$. 我们随后给这个集合配备产品拓扑的弱收敛拓扑在预估器和通常的离散拓扑在信息变量上。这样，我们不将信念状态映射到集合 $\mathring{Z}_N^{\pi}$ 中的最近元素，而是将信念状态映射到元素 $I_t^N$ 中具有匹配信息变量 $I_t^N$. 其他，真实信念 $(\pi_t^{-N}, I_t^N)$ 总是被映射/量化为 $(\pi^*, I_t^N)$. 注意到这种映射可以不通过预估器 $\pi_t^{-N}$ 的知识或计算进行，只要我们有 $I_t^N$ 可用，因此，设计者不需要计算或跟踪原始信念状态，但只需跟踪 $I_t^N$ 的窗口信息变量 $I_t^N = Y_{[0,N]}, U_{[0,N-1]}$.

我们将看到在第6章中，使用 $I_t^N$ 构建的近似Q学习算法将收敛于第5.4节中构建的近似模型。

在总结中，我们提出的第一种方法在第5.3节中提供更好的近似性能。然而，我们提出在第5.4节中的第二种方法提供了一个相当重要的实际优势，但代价是性能。

5.6 数值研究

在这一节中，我们呈现了证明结果的数值研究。

我们使用的例子是一个机器维修问题。在这个模型中，我们有 $X, Y, U = \{0, 1\}$ 与

$$x_t = \begin{cases} 1 & \text{machine is working at time } t \\ 0 & \text{machine is not working at time } t \\ \end{cases}$$

$$u_t = \begin{cases} 1 & \text{machine is being repaired at time } t \\ 0 & \text{machine is not being repaired at time } t \\ \end{cases}$$
The one stage cost function is given by

\[
    c(x, u) = \begin{cases} 
        R + E & x = 0, u = 1 \\
        E & x = 0, u = 0 \\
        0 & x = 1, u = 0 \\
        R & x = 1, u = 1 
    \end{cases}
\]

where \( R \) is the cost of repair and \( E \) is the cost incurred by a broken machine.

We study the example with discount factor \( \beta = 0.8 \). The transition probabilities are chosen as follows:

\[
    Pr(x_1 = 0 | x_0 = 0, u_0 = 0) = 0.9, \quad Pr(x_1 = 0 | x_0 = 0, u_0 = 1) = 0.6 \\
    Pr(x_1 = 0 | x_0 = 1, u_0 = 0) = 0.4, \quad Pr(x_1 = 0 | x_0 = 1, u_0 = 1) = 0.1
\]

For the measurement channel:

\[
    Pr(y = 0 | x = 0) = 0.7, \quad Pr(y = 1 | x = 1) = 0.7.
\]

For the cost function, we choose \( R = 3 \), and \( E = 1 \).

Finally, we choose the starting distribution as \( \bar{\pi}(\cdot) = 0.42\delta_0(\cdot) + 0.58\delta_1(\cdot) \).

We first compare the approximate policies calculated using the nearest neighbor method from Section 5.3 and the history matching method from Section 5.4. We plot the following terms for the window size ranging from 0 to 5

\[
    \text{Performance Loss: } E^\gamma_0[|J_\beta(Z_5, \gamma^N_i) - J^*_\beta(Z_5)|Y_{[0,5]}, U_{[0,4]}], \quad \text{for } i \in \{1, 2\}
\]
where $\gamma_0$ is some known initial policy for the first 5 time steps. $\gamma_N^1$ denotes the approximate policy for the nearest neighbor method from Section 5.3 and $\gamma_N^2$ denotes the approximate policy for the history matching method from Section 5.4.

To get the errors, we simply subtract the cost values from their minimum (largest window) which serves as an approximation of the optimal value function.

![Figure 5.1: Approximation error in memory size for different methods.](image)

We can see that the approximate policy calculated using the approach in Section 5.3 performs slightly better, but both methods tend to perform near optimally with increasing window size.

We now plot the performance of the policy from Section 5.3 seperately compared to the $\rho_{BL}$ filter stability term

$$\sup_{\gamma} \sup_{\pi} E_\pi^\gamma \left[ \rho_{BL} \left( P^\pi(X_N \in \cdot | Y_{[0,N]}), P^{\hat{\pi}}(X_N \in \cdot | Y_{[0,N]}) \right) \right]$$

and $\alpha^N$ where the Dobrushin constant is $\alpha = (2 - \delta(Q)) \times (1 - \delta(T))$. Recall that in Theorem 5.3.1 and Theorem 5.3.2 we proved that $\alpha^N$ dominates the expected filter stability term which in turn dominates the approximation error.
5.6. NUMERICAL STUDY

Figure 5.2: Approximation error, BL filter stability and Dobrushin coefficient in memory size for the first method.

Notice that the graph verifies what have proved in Section 5.3.

We now compare the error for the second method presented in Section 5.4 and the total variation filter stability term

$$\sup_{\gamma} \sup_{\pi} E^\gamma \left[ \| P^\pi (X_N \in \cdot | Y_{[0,N]}) - \hat{P}^\pi (X_N \in \cdot | Y_{[0,N]}) \|_{TV} \right]$$

which coincides with the $\rho_{BL}$ metric since we use a finite space $X$.

Figure 5.3: Approximation error, TV filter stability and Dobrushin coefficient in memory size for the second method.
Notice that this graph verifies Theorem 5.4.2.

Finally, we plot everything together for an overall comparison:

![Overall comparison graph](image)

Figure 5.4: Overall comparison

5.7 Proofs of Technical Results

5.7.1 Proof of Theorem 5.3.1

Before presenting our proof program and the series of technical results needed, we introduce some notation.

We denote the loss function due to the quantization by $L(z)$, i.e.

$$L(z) = \rho_{BL}(z, F(z))$$

with $F$ defined in (5.4).

In the following, to specify the probability measures according to which the expectations are taken, we use the following notation; for the kernel $\eta^N$ and a policy $\gamma$, we use $E_N^\gamma$ and for the kernel $\eta$ and a policy $\gamma$, we will use $E^\gamma$.

The last notation we introduce is the following: Recall that we denote the optimal cost
for the finite model by $J^N_\beta$ and the optimal policy by $\gamma^*_N$. These are defined on a finite set $\mathcal{Z}^N_\beta$, however, we can always extend them over the original state space $\mathcal{Z}$ so that they are constant over the quantization bins. We denote the extended versions by $\tilde{J}^N_\beta$ and $\tilde{\gamma}^*_N$.

Now, we introduce our value iteration approach for the original model and the finite model. We write for any $k < \infty$

$$v_{k+1}(z) = \min_u \left( \tilde{c}(z, u) + \beta \int v_k(z_1)\eta(dz_1|z, u) \right) \quad \forall z \in \mathcal{Z},$$

$$v^N_{k+1}(z) = \min_u \left( c^N(z, u) + \beta \sum_{z_1} v^N_k(z_1)\eta^N(z_1|z, u) \right) \quad \forall z \in \mathcal{Z}^N_\beta.$$ 

where $v_0, v^N_0 \equiv 0$. It is well known that this operator is a contraction under either model and hence the value functions converge to the optimal expected discounted cost. In particular, we have that

$$|J^N_\beta(z) - v^N_k(z)| \leq \|\tilde{c}\|_\infty \frac{\beta^k}{1-\beta}, \quad |J^*_\beta(z) - v_k(z)| \leq \|\tilde{c}\|_\infty \frac{\beta^k}{1-\beta}.$$ 

Notice that if we extend $v^N_{k+1}(z)$ and $c^N(z, u)$ over all of the state space then $\tilde{v}^N_{k+1}(z)$ and $\tilde{c}^N(z, u)$ becomes constant over the quantization bins. Then, the extended value functions for the finite model can be also obtained with

$$\tilde{v}^N_{k+1}(z) = \min_u \left( \tilde{c}(F(z), u) + \beta \int \tilde{v}^N_k(z_1)\eta(dz_1|F(z), u) \right) \quad \forall z \in \mathcal{Z}. \quad (5.18)$$

To see this, observe the following:

$$\tilde{v}^N_{k+1}(z) = \min_u \left( c(F(z), u) + \beta \int \tilde{v}^N_k(z_1)\eta(dz_1|F(z), u) \right)$$
5.7. PROOFS OF TECHNICAL RESULTS

\[
= \min_u \left( \hat{c}(F(z), u) + \beta \sum_{i=1}^{\lfloor Z_n^i \rfloor} \int_{Z_i^N} \tilde{v}_k^N(z_1) \eta(dz_1|F(z), u) \right)
\]

\[
= \min_u \left( \hat{c}(F(z), u) + \beta \sum_{i=1}^{\lfloor Z_n^i \rfloor} \int_{Z_i^N} \tilde{v}_k^N(z_1^i) \eta(dy|F(z), u) \right)
\]

\[
= \min_u \left( \hat{c}(F(z), u) + \beta \sum_{i=1}^{\lfloor Z_n^i \rfloor} \tilde{v}_k^N(z_1^i) \eta(Z_i^N|F(z), u) \right)
\]

where \( Z_i^n \) denotes the \( i \)th quantization bin and \( \tilde{v}_k^N(z_1^i) \) denotes the value of \( \hat{v}_k^N \) at that quantization bin and it is constant over the bin.

For the proof the goal is to bound the term \( |J_\beta(\tilde{\gamma}_N^*, z) - J_\beta(\gamma^*, z)| \). We will see in the following that bounding this term is related to studying the term \( |\tilde{J}_\beta^N(\tilde{\gamma}_N^*, z) - J_\beta(\gamma^*, z)| \) (the value function approximation error). Therefore, in what follows, we first analyze \( |\tilde{J}_\beta^N(\tilde{\gamma}_N^*, z) - J_\beta(\gamma^*, z)| \) and observe that it can be bounded using the expected loss terms \( E[L(Z_t)] \). Finally, we will observe that upper bounds on these expressions can be written through the filter stability term

\[
\sup_{\pi \in P(X)} \sup_{\gamma \in \Gamma} E^\gamma \left[ \rho_{BL} \left( \tilde{\pi}_N(X_N \in [0,N], \gamma(Y_{[0,N-1]})), \tilde{\pi}_N(X_N \in [0,N], \gamma(Y_{[0,N-1]})) \right) \right].
\]

Next, we present some supporting technical results.

**Lemma 5.7.1.** Under Assumption 5.3.1 we have

\[
|\tilde{J}_\beta^N(\tilde{\gamma}_N^*, z) - J_\beta(\gamma^*, z)|
\leq \lim_{k \to \infty} \sup_{\gamma \in \Gamma} \left( \alpha \tilde{c} \sum_{t=0}^{k-1} \beta^t \left( E_z^\gamma [L(Z_t)] + 3 \alpha_z \sum_{m=0}^{t-1} (4 \alpha_z + 1)^m E_z^\gamma [L(Z_{t-m-1})] \right) + \alpha \sum_{t=0}^{k-1} \beta^{t+1} \|v_{k-t-1}\|_{BL} \left( E_z^\gamma [L(Z_t)] + 3 \alpha_z \sum_{m=0}^{t-1} (4 \alpha_z + 1)^m E_z^\gamma [L(Z_{t-m-1})] \right) \right).
\]
where $v_k$ denotes the value function at a time step $k$ that is produced by the value iteration with $v_0 \equiv 0$.

**Proof.** We follow a value iteration approach to write for any $k < \infty$

\[
v_{k+1}(z) = \min_u \left( \tilde{c}(z, u) + \beta \int v_k(y) \eta(dy|z, u) \right) \quad \forall z \in \mathcal{Z},
\]

\[
\tilde{v}_{k+1}^N(z) = \min_u \left( \tilde{c}(F(z), u) + \beta \int \tilde{v}_k^N(y) \eta(dy|F(z), u) \right) \quad \forall z \in \mathcal{Z}.
\]

where $v_0, \tilde{v}_0^N \equiv 0$.

We then write

\[
|\tilde{J}_{\beta}^N(z) - J_{\beta}^*(z)| \leq |\tilde{J}_{\beta}^N(z) - \tilde{v}_k^N(z)| + |\tilde{v}_k^N(z) - v_k(z)| + |v_k(z) - J_{\beta}^*(z)| \tag{5.19}
\]

For the first and the last term, using the fact that the dynamic programming operator is a contraction, we have that

\[
|\tilde{J}_{\beta}^N(z) - \tilde{v}_k^N(z)| \leq \|\tilde{c}\|_{\infty} \frac{\beta_0^k}{1 - \beta}, \quad |J_{\beta}^*(z) - v_k(z)| \leq \|\tilde{c}\|_{\infty} \frac{\beta_0^k}{1 - \beta}.
\]

Now, we focus on the second term $|\tilde{v}_k^N(z) - v_k(z)|$.

**Step 1:** We show in Section 5.7.2 that for all $k \geq 1$,

\[
|\tilde{v}_k^N(z) - v_k(z)| \leq \sup_{\gamma \in \Gamma} \left( \alpha_\epsilon \sum_{t=0}^{k-1} \beta^t E_{z,N}^\gamma [L(Z_t)] + \sum_{t=0}^{k-1} \beta^{t+1} \|v_{k-t-1}\|_{BL} E_{z,N}^\gamma [L(Z_t)] \alpha_\epsilon \right)
\]

where $L(z)$ is the loss function due to the quantization, i.e. $L(z) = \rho_{BL}(z, \hat{z})$ where $\hat{z}$ is representative state $z$ belongs to.
Step 2: We show in Section 5.7.3 that

$$E_{\gamma}^{\gamma}[L(Z_t)] \leq E_{\gamma}^{\gamma}[L(Z_t)] + 3\alpha_z \sum_{m=0}^{t-1} (4\alpha_z + 1)^m E_{\gamma}^{\gamma}[L(Z_{t-m-1})].$$

Step 3: Combining our results, we have that

$$|\tilde{v}_k^N(z) - v_k(z)|$$

$$\leq \sup_{\gamma} \left( \alpha_c \sum_{t=0}^{k-1} \beta^t E_{\gamma}^{\gamma}[L(Z_t)] + \sum_{t=0}^{k-1} \beta^{t+1} ||v_{k-t-1}||BL E_{\gamma}^{\gamma}[L(Z_t)] \alpha_z \right)$$

$$\leq \sup_{\gamma} \left( \alpha_c \sum_{t=0}^{k-1} \beta^t \left( E_{\gamma}^{\gamma}[L(Z_t)] + 3\alpha_z \sum_{m=0}^{t-1} (4\alpha_z + 1)^m E_{\gamma}^{\gamma}[L(Z_{t-m-1})] \right) 
 + \alpha_z \sum_{t=0}^{k-1} \beta^{t+1} ||v_{k-t-1}||BL \left( E_{\gamma}^{\gamma}[L(Z_t)] + 3\alpha_z \sum_{m=0}^{t-1} (4\alpha_z + 1)^m E_{\gamma}^{\gamma}[L(Z_{t-m-1})] \right) \right).$$

(5.20)

Hence, using (5.19) and (5.20), we write

$$|\tilde{J}^N_{\beta}(z) - J^*_\beta(z)| \leq \lim_{k \to \infty} |\tilde{v}_k^N(z) - v_k(z)|$$

$$\leq \lim_{k \to \infty} \sup_{\gamma} \left( \alpha_c \sum_{t=0}^{k-1} \beta^t \left( E_{\gamma}^{\gamma}[L(Z_t)] + 3\alpha_z \sum_{m=0}^{t-1} (4\alpha_z + 1)^m E_{\gamma}^{\gamma}[L(Z_{t-m-1})] \right) 
 + \alpha_z \sum_{t=0}^{k-1} \beta^{t+1} ||v_{k-t-1}||BL \left( E_{\gamma}^{\gamma}[L(Z_t)] + 3\alpha_z \sum_{m=0}^{t-1} (4\alpha_z + 1)^m E_{\gamma}^{\gamma}[L(Z_{t-m-1})] \right) \right).$$

The next result, gives a bound on the loss function occurring from the quantization (on $L(z)$) and relates the bound to the filter stability problem.

**Lemma 5.7.2.** For $Z_0 = P_{\pi_0}(X_N \in \cdot|Y_{[0,N]}, \gamma(Y_{[0,N-1]}))$, where $\pi_0 \in \mathcal{P}(X)$ is the prior
5.7. PROOFS OF TECHNICAL RESULTS

distribution of $X_0$, we have that for any $N < t < \infty$

$$\sup_{\gamma \in \Gamma} E_{\pi_0}^\gamma \left[ E_{Z_0}^\gamma \left[ L(Z_t) \right] \middle| Y_{[0,N]}, \gamma(Y_{[0,N-1]}) \right]$$

$$\leq \sup_{\pi \in \mathcal{P}(X)} \mathbb{E}_\pi \left[ \rho_{BL} \left( P^\pi(X_N \in \cdot | Y_{[0,N]}, \gamma(Y_{[0,N-1]})), P^\pi(X_N \in \cdot | Y_{[0,N]}), \gamma(Y_{[0,N-1]}) \right) \right].$$

**Proof.** We first note that, since the quantizer is a nearest-neighbor quantization, the quantization error is almost surely upper bounded by

$$\rho_{BL} \left( P^{\pi_{t-1}}(X_{N+t} \in \cdot | Y_{[t,t+N]}, U_{[t,t+N-1]}), P^\pi(X_{N+t} \in \cdot | Y_{[t,t+N]}, U_{[t,t+N-1]}) \right).$$

Using this and the law of total expectation:

$$E_{\pi_0}^\gamma \left[ E_{Z_0}^\gamma \left[ L(Z_t) \middle| Y_{[0,N]}, \gamma(Y_{[0,N-1]}) \right] \right]$$

$$\leq \sup_{\gamma \in \Gamma} E_{\pi_{t-1}}^\gamma \left[ \rho_{BL} \left( P^{\pi_{t-1}}(X_{N+t} \in \cdot | Y_{[t,t+N]}, U_{[t,t+N-1]}) \right) \right] \left[ (U,Y)_{[0,t-1]} \right]$$

$$\leq \sup_{\pi \in \mathcal{P}(X)} \mathbb{E}_\pi \left[ \rho_{BL} \left( P^\pi(X_N \in \cdot | Y_{[0,N]}, U_{[0,N-1]}), P^\pi(X_N \in \cdot | Y_{[0,N]}, U_{[0,N-1]}) \right) \right]$$

where $$(U,Y)_{[0,t-1]} = U_{[0,t-1]}, Y_{[0,t-1]}.$$

**Lemma 5.7.3.** Under Assumption [5.3.1] we have that

$$\|v_k\|_{BL} \leq C \frac{1}{1 - \beta} \left( \|\tilde{c}\|_\infty + \sum_{t=0}^{k-1} (\beta \alpha_2)^t \right).$$

In particular, we have that for all $k$

$$\|v_k\|_{BL} \leq \frac{1}{1 - \beta \alpha_2} \left( \|\tilde{c}\|_\infty + \alpha \right).$$
5.7. PROOFS OF TECHNICAL RESULTS

Proof. It is easy to see that \( \|v_k\|_\infty \leq \|\tilde{c}\|_\infty \sum_{t=0}^{k-1} \beta^t = \|\tilde{c}\|_\infty \frac{1 - \beta^k}{1 - \beta} \). Then, we use an inductive approach and assume the claim holds for \( k \) and analyze the term \( k + 1 \)

\[
\|v_{k+1}\|_{BL} = \|v_{k+1}\|_\infty + \sup_{x \neq y} \frac{|v_{k+1}(x) - v_{k+1}(y)|}{\rho_{BL}(x, y)}
\]

For the second term, we have

\[
|v_{k+1}(x) - v_{k+1}(y)| \leq \sup_u \left( |\tilde{c}(x, u) - \tilde{c}(y, u)| + \beta \left| \int v_k(z) \eta(dz|x, u) - \int v_k(z) \eta(dz|y, u) \right| \right) \\
\leq \alpha \varepsilon \rho_{BL}(x, y) + \beta \|v_k\|_{BL} \alpha z \rho_{BL}(x, y).
\]

Hence, using the induction hypothesis, we have that

\[
\|v_{k+1}\|_{BL} \leq \|v_{k+1}\|_\infty + \alpha \varepsilon + \beta \alpha z \|v_k\|_{BL} \\
\leq \|\tilde{c}\|_\infty \frac{1 - \beta^{k+1}}{1 - \beta} + \alpha \varepsilon + \beta \alpha z \left( \alpha \varepsilon \sum_{t=0}^{k-1} (\beta \alpha z)^t + \|\tilde{c}\|_\infty \sum_{t=0}^{k-1} (\beta \alpha z)^t \frac{1 - \beta^{k-t}}{1 - \beta} \right) \\
= \alpha \varepsilon \sum_{t'=0}^{k} (\beta \alpha z)^{t'} + \|\tilde{c}\|_\infty \sum_{t'=0}^{k} (\beta \alpha z)^{t'} \frac{1 - \beta^{k-t'}}{1 - \beta}, \ t' = t + 1,
\]

which concludes the induction argument. Therefore, we can write

\[
\|v_k\|_{BL} \leq \alpha \varepsilon \sum_{t=0}^{k-1} (\beta \alpha z)^t + \|\tilde{c}\|_\infty \sum_{t=0}^{k-1} (\beta \alpha z)^t \frac{1 - \beta^{k-t}}{1 - \beta} = \left( \alpha \varepsilon + \|\tilde{c}\|_\infty \frac{1 - (\beta \alpha z)^k}{1 - \beta \alpha z} \right) \frac{1 - \beta^{k-t}}{1 - \beta} + \|\tilde{c}\|_\infty \frac{\beta^{k-t}}{1 - \beta} \frac{1 - \beta^{k-t}}{1 - \beta} \frac{1 - \beta^{k-t}}{1 - \beta} \frac{1 - \beta^{k-t}}{1 - \beta} \frac{1 - \beta^{k-t}}{1 - \beta} \frac{1 - \beta^{k-t}}{1 - \beta}
\]

Now, we can prove Theorem 5.3.1.

Proof. of Theorem 5.3.1 Consider the following dynamic programming operators, \( \hat{T}_n, T_n : \mathcal{B}_b(\mathcal{Z}) \to \mathcal{B}_b(\mathcal{Z}) \) (\( \mathcal{B}_b(\mathcal{Z}) \) denotes the set of measurable and bounded functions on \( \mathcal{Z} \)) such
that for any $f \in \mathcal{B}_b(\mathcal{Z})$ and for any $z \in \mathcal{Z}$

$$(T_n(f))(z) := c(z, \tilde{\gamma}_N^*(z)) + \beta \int f(y) \eta(dy|z, \tilde{\gamma}_N^*(z))$$

$$(\hat{T}_n(f))(z) := c(F(z), \tilde{\gamma}_N^*(z)) + \beta \int f(y) \eta(dy|F(z), \tilde{\gamma}_N^*(z))$$

where $\tilde{\gamma}_N^*$ denotes the extension of the optimal policy for the finite state space model to the belief space $\mathcal{Z}$, i.e. it is constant over the quantization bins, furthermore $F(z)$ is the nearest neighbor quantization map such that it maps $z$ to the closest point from the finite state space $\mathcal{Z}_N$ (recall that $\rho_{BL}$ is used to metrize the belief space).

The optimal cost for the finite model, $J_N^\beta$ is only defined on a finite set, we denote by $\tilde{J}_N^\beta$ the extension of the optimal cost for the finite model to the belief space $\mathcal{P}(\mathcal{X})$ by assigning the same values over the quantization bins, i.e. it is a piece-wise constant function over the quantization bins. Note that, we have

$$J_\beta(\tilde{\gamma}_N^*, z) = T_n(J_\beta(\tilde{\gamma}_N^*, z)),$$

$$\tilde{J}_N^\beta(\tilde{\gamma}_N^*, z) = \hat{T}_n(\tilde{J}_N^\beta(\tilde{\gamma}_N^*, z)).$$

Using these equalities, we write

$$J_\beta(\tilde{\gamma}_N^*, z) - J_\beta(\gamma^*, z) \leq |T_n(J_\beta(\tilde{\gamma}_N^*, z)) - T_n(J_\beta(\gamma^*, z))|$$

$$+ |T_n(J_\beta(\gamma^*, z)) - \hat{T}_n(J_\beta(\gamma^*, z))|$$

$$+ |\hat{T}_n(J_\beta(\gamma^*, z)) - \hat{T}_n(\tilde{J}_N^\beta(\tilde{\gamma}_N^*, z))|$$

$$+ |\tilde{J}_N^\beta(\gamma^*, z) - J_\beta(\gamma^*, z)|$$

$$\leq \beta \int |J_\beta(\tilde{\gamma}_N^*, z_1) - J_\beta(\gamma^*, z_1)| \eta(dz_1|z, \tilde{\gamma}_N^*(z))$$
\[ + |c(z, \tilde{\gamma}_N^*(z)) - c(F(z), \tilde{\gamma}_N^*(z))| \]
\[ + \beta | \int J_\beta(\gamma^*, z_1)\eta(dz_1|z, \tilde{\gamma}_N^*(z)) - \int J_\beta(\gamma^*, z_1)\eta(dz_1|F(z), \tilde{\gamma}_N^*(z))| \]
\[ + \beta | \int J_\beta^N(\tilde{\gamma}_N^*, z_1) - J_\beta(\gamma^*, z_1)|\eta(dz_1|F(z), \tilde{\gamma}_N^*(z)) \]
\[ + |\tilde{J}_\beta^N(\tilde{\gamma}_N^*, z) - J_\beta(\gamma^*, z)| \]
\[ \leq \beta \sup_\gamma \left( E^\gamma_z[|J_\beta(\tilde{\gamma}_N^*, Z_1) - J_\beta(\gamma^*, Z_1)|] + \alpha_c L(z) + \beta \alpha_z L(z)\|J_\beta^N\|_{BL} \right. \]
\[ \left. + \beta E^\gamma_{z,N}[|\tilde{J}_\beta^N(\tilde{\gamma}_N^*, Z_1) - J_\beta(\gamma^*, Z_1)| + |\tilde{J}_\beta^N(\tilde{\gamma}_N^*, z) - J_\beta(\gamma^*, z)|] \right) \quad (5.21) \]

where \( E^\gamma_{z,N} \) denotes the expectation with respect the the kernel \( \eta(dy|F(z), \gamma(z)) \) when the initial point is \( z \).

Now, using Lemma 5.7.1, we can write that:

\[ |\tilde{J}_\beta^N(\tilde{\gamma}_N^*, z) - J_\beta(\gamma^*, z)| \]
\[ \leq \lim_{k \to \infty} \sup_\gamma \left( \alpha_c \sum_{t=0}^{k-1} \beta^t \left( E^\gamma_z[L(Z_t)] + 3 \alpha_z \sum_{m=0}^{t-1} (4\alpha_z + 1)^m E^\gamma_z[L(Z_{t-m-1})] \right) \right. \]
\[ \left. + \alpha_z \sum_{t=0}^{k-1} \beta^{t+1} \|v_{k-t-1}\|_{BL} \left( E^\gamma_z[L(Z_t)] + 3 \alpha_z \sum_{m=0}^{t-1} (4\alpha_z + 1)^m E^\gamma_z[L(Z_{t-m-1})] \right) \right) \]
\[ := g(z). \quad (5.22) \]

We first introduce the following notation along with the above notation (5.22):

\[ f(z) := |J_\beta(\tilde{\gamma}_N^*, z) - J_\beta(\gamma^*, z)|. \]
Notice that with the new notation, we can rewrite the bound on (5.21) as:

\[
f(z) \leq \beta \sup_{\gamma} E_{z,0}^\gamma [f(Z_1)] + \alpha_c L(z) + \beta \alpha_z L(z) \|J^*_\beta\|_{BL} + \beta \gamma E_{z,N}^\gamma [g(Z_1)] + g(z)
\]

where we used the fact that

\[
E_{z,N}^\gamma[g(Z_1)] - E_z^\gamma[g(Z_1)] = \int g(z_1) \eta(dz_1 | F(z), \gamma(z)) - \int g(z_1) \eta(dz_1 | z, \gamma(z)) \leq \|g\|_{BL} L(z).
\]

Using the same bound on \( f(Z_1) \) one can also write that for any initial point \( z_0 \):

\[
f(z_0) \leq \sup_{\gamma \in \Gamma} \beta^2 E_{z_0}^\gamma [f(Z_2)] + \alpha_c \sum_{t=0}^{1} \beta^t E_{z_0}^\gamma [L(Z_t)] + \alpha_z \|J^*_\beta\|_{BL} \sum_{t=0}^{1} \beta^{t+1} E_{z_0}^\gamma [L(Z_t)]
\]

\[
+ \sum_{t=0}^{1} \beta^{t+1} E_{z_0}^\gamma [g(Z_t)] + \|g\|_{BL} \sum_{t=0}^{1} \beta^{t+1} E_{z_0}^\gamma [L(Z_t)] + \sum_{t=0}^{1} \beta^{t+1} E_{z_0}^\gamma [g(Z_t)].
\]

In general, for any \( k < \infty \), we can write

\[
f(z_0) \leq \sup_{\gamma \in \Gamma} \left( \beta^k E_{z_0}^\gamma [f(Z_k)] + \alpha_c \sum_{t=0}^{k-1} \beta^t E_{z_0}^\gamma [L(Z_t)] + \alpha_z \|J^*_\beta\|_{BL} \sum_{t=0}^{k-1} \beta^{t+1} E_{z_0}^\gamma [L(Z_t)]
\]

\[
+ \sum_{t=0}^{k-1} \beta^{t+1} E_{z_0}^\gamma [g(Z_t)] + \|g\|_{BL} \sum_{t=0}^{k-1} \beta^{t+1} E_{z_0}^\gamma [L(Z_t)] + \sum_{t=0}^{k-1} \beta^{t+1} E_{z_0}^\gamma [g(Z_t)] \right). \quad (5.23)
\]

Recall that our main goal is to bound \( E_{\pi_0}^\cdot [f(Z_0)|Y_{[0,N]}, \gamma(Y_{[0,N-1]})] \) where \( Z_0 = P_{\pi_0}(X_N \in \cdot|Y_{[0,N]}, \gamma(Y_{[0,N-1]})). \) To this end, first notice that using Lemma 5.7.2, we have for any \( t \)

\[
E_{\pi_0}^\cdot [E_{Z_0}^\gamma [L(Z_t)]|Y_{[0,N]}, \gamma(Y_{[0,N-1]})]
\]
\[ \leq \sup_{\pi \in \mathcal{P}(X)} \sup_{\gamma \in \Gamma} E^{\pi, \gamma} \left[ \rho_{BL} \left( P^{\pi}(X_N \in \cdot | Y_{[0,N]}), P^{\pi}(X \in \cdot | Y_{[0,N]}) \right) \right] := B. \quad (5.24) \]

Using this bound and Lemma 5.7.3 for \( \| v_{k-t-1} \|_{BL} \), we can write that

\[
E \left[ E^{\gamma}_{Z_0} [g(Z_t)] | Y_{[0,N]}, \gamma(Y_{[0,N-1]}) \right] \\
\leq \lim_{k \to \infty} \left( \alpha_c \sum_{t=0}^{k-1} \beta^t \left( B + 3\alpha_z \sum_{m=0}^{t-1} B(4\alpha_z + 1)^m \right) + \alpha_z \sum_{t=0}^{k-1} \beta^{t+1} \| v_{k-t-1} \|_{BL} \left( B + 3\alpha_z \sum_{m=0}^{t-1} B(4\alpha_z + 1)^m \right) \right) \\
\leq B \left( \alpha_c + \alpha_z \| J^*_\beta \|_{BL} \right) \lim_{k \to \infty} \sum_{t=0}^{k-1} \beta^t \left( 1 + 3\alpha_z \sum_{m=0}^{t-1} (4\alpha_z + 1)^m \right) \\
\leq B \left( \alpha_c + \alpha_z \| J^*_\beta \|_{BL} \right) \lim_{k \to \infty} \sum_{t=0}^{k-1} \beta^t (4\alpha_z + 1)^t \\
= BK_0(\beta, \alpha_z, \alpha_{\tilde{c}}, \| \tilde{c} \|_{\infty}) \quad (5.25) \]

where

\[
K_0(\beta, \alpha_z, \alpha_{\tilde{c}}, \| \tilde{c} \|_{\infty}) = \left( \alpha_c + \beta \alpha_z \| J^*_\beta \|_{BL} \right) \frac{1}{1 - \beta(4\alpha_z + 1)}. \]

In Section 5.7.4 we show that

\[
\| g \|_{BL} \leq K_0(\beta, \alpha_z, \alpha_{\tilde{c}}, \| \tilde{c} \|_{\infty}) \frac{3}{1 - \alpha_z}. \]
Going back to (5.23), using (5.24) and (5.25) and taking the limit \( k \to \infty \):

\[
E[f(Z_0)|Y_{[0,N]}, \gamma(Y_{[0,N-1]})] \leq \frac{B\alpha_c}{1-\beta} + \frac{\beta\alpha_z \|J^*_\beta\|_{BL} B}{1-\beta} + \frac{\beta K_0(\beta, \alpha_z, \alpha_{\tilde{c}}, \|\tilde{c}\|_\infty)B}{1-\beta} \\
+ \frac{\beta K_0(\beta, \alpha_z, \alpha_{\tilde{c}}, \|\tilde{c}\|_\infty)B}{1-\beta} + \frac{K_0(\beta, \alpha_z, \alpha_{\tilde{c}}, \|\tilde{c}\|_\infty)B}{1-\beta}.
\]

Thus, we can write that

\[
E[f(Z_0)|Y_{[0,N]}, \gamma(Y_{[0,N-1]})] \leq K(\beta, \alpha_z, \alpha_{\tilde{c}}, \|\tilde{c}\|_\infty) \sup_{\pi \in \mathcal{P}(X)} \sup_{\gamma \in \Gamma} E^{\pi,\gamma} \left[ \rho_{BL} \left( P^\pi(X_N \in \cdot|Y_{[0,N]}), P^{\hat{\pi}}(X_N \in \cdot|Y_{[0,N]}) \right) \right],
\]

where

\[
K(\beta, \alpha_z, \alpha_{\tilde{c}}, \|\tilde{c}\|_\infty) = \frac{\alpha_c + \beta \alpha_z \|J^*_\beta\|_{BL} + \beta K_0(\beta, \alpha_z, \alpha_{\tilde{c}}, \|\tilde{c}\|_\infty)(1 + \frac{3}{1-\alpha_z}) + K_0(\beta, \alpha_z, \alpha_{\tilde{c}}, \|\tilde{c}\|_\infty)}{1-\beta}
\]

and using Lemma 5.7.6 (see Section 5.7.5)

\[
\|J^*_\beta\|_{BL} \leq \frac{1}{1-\beta \alpha_z} \left( \frac{\|\tilde{c}\|_\infty}{1-\beta} + \alpha_{\tilde{c}} \right).
\]

\[\square\]

5.7.2 Proof of Step 1 in the Proof of Lemma 5.7.1

We claim that

\[
|\tilde{v}_k^N(z) - v_k(z)| \leq \sup_{\gamma \in \Gamma} \left( \sum_{t=0}^{k-1} \beta^t E_{z,\gamma}^N[L(Z_t)] + \sum_{t=0}^{k-1} \beta^{t+1} \|v_{k-t-1}\|_{BL} E_{z,\gamma}^N[L(Z_t)] \alpha_z \right)
\]
where $L(z)$ is the loss function due to the quantization, i.e. $L(z) = \rho_{BL}(z, F(z))$ with $F$ defined in (5.4).

We prove the claim using an inductive approach: for $k = 1$ we have (noting $v_0, \tilde{v}_0^N \equiv 0$)

$$
\tilde{v}_1^N(z) = \min_u \left( \tilde{c}(F(z), u) + \beta \int \tilde{v}_0^N(y) \eta(dy|F(z), u) \right) = \min_u \tilde{c}(F(z), u)
$$

$$
v_1(z) = \min_u \left( \tilde{c}(z, u) + \beta \int v_0(y) \eta(dy|z, u) \right) = \min_u \tilde{c}(z, u).
$$

Note that under the stated assumptions the measurable selection conditions hold, and the minimum can be achieved using a policy $\gamma$ for the original model and a policy $\gamma^N$ for the finite model which defined only on a finite set. By extending the finite model policy $\gamma^N$ over all state space $\mathcal{P}(X)$, we can write that

$$
|\tilde{v}_1^N(z) - v_1(z)| \leq \sup_{\gamma} \left( |\tilde{c}(F(z), \gamma(z)) - \tilde{c}(z, \gamma(z))| + |\tilde{c}(z, \gamma^N(z)) - \tilde{c}(F(z), \gamma^N(z))| \right)
$$

$$
\leq \alpha \varepsilon L(z)
$$

which completes the proof for $k = 1$. Now, we assume that the claim holds for $k$ and analyze the step $k + 1$. Similar to $k = 1$ case, we can write

$$
|\tilde{v}_{k+1}^N(z) - v_{k+1}(z)|
$$

$$
\leq \sup_{\gamma} \left( |\tilde{c}(F(z), \gamma(z)) - \tilde{c}(z, \gamma(z))| + \beta \int |\tilde{v}_k^N(y) - v_k(y)| \eta(dy|F(z), \gamma(z)) 
\right)
$$

$$
\leq \sup_{\gamma} \left( |\tilde{c}(F(z), \gamma(z)) - \tilde{c}(z, \gamma(z))| + \beta \int |\tilde{v}_k^N(y) - v_k(y)| \eta(dy|F(z), \gamma(z)) 
\right)
$$

$$
+ \beta \left| \int v_k(y) \eta(dy|F(z), \gamma(z)) - \int v_k(y) \eta(dy|z, \gamma(z)) \right|
$$
5.7. PROOFS OF TECHNICAL RESULTS

\[ \leq \alpha \bar{c}L(z) + \beta \sup_{\gamma} \left( \int \left( \alpha \bar{c} \sum_{t=0}^{k-1} \beta t E_{y,\gamma}^N [L(Z_t)] + \sum_{t=0}^{k-1} \beta^{t+1} \| v_{k-t-1} \|_{BL} E_{y,\gamma}^N [L(Z_t)] \right) \eta(dy | F(z), \gamma(z)) \right) \]

\[ + \beta \| v_k \|_{BL} \alpha \bar{c}L(z) \]

\[ \leq \sup_{\gamma} \left( \alpha \bar{c}L(z) + \alpha \bar{c} \sum_{t=0}^{k-1} \beta^{t+1} E_{z,\gamma} [L(Z_{t+1})] + \beta \| v_k \|_{BL} \alpha \bar{c}L(z) + \sum_{t=0}^{k-1} \beta^{t+2} \| v_{k-t-1} \|_{BL} E_{z,\gamma}^N [L(Z_{t+1})] \right) \]

\[ = \sup_{\gamma} \left( \alpha \bar{c} \sum_{t'=0}^{k} \beta^{t'} E_{z,\gamma} [L(Z_{t'})] + \sum_{t'=0}^{k} \beta^{t'+1} \| v_{k-t'-1} \|_{BL} E_{z,\gamma}^N [L(Z_{t'})] \right), \quad (t' = t + 1). \]

For the last two steps, note that \( E_{y,\gamma}^N [L(Z_t)] \) denotes the expected loss at time \( t \) when the initial state \( Z_0 = y \), thus using the iterative expectation we have

\[ \int E_{y,\gamma}^N [L(Z_t)] \eta(dy | F(z), \gamma(z)) = E_{z,\gamma}^N [L(Z_{t+1})]. \]

Hence, we have proved that for all \( k \geq 1 \)

\[ |v_k^N (z) - v_k (z)| \leq \sup_{\gamma} \left( \alpha \bar{c} \sum_{t=0}^{k-1} \beta^{t} E_{z,\gamma}^N [L(Z_t)] + \sum_{t=0}^{k-1} \beta^{t+1} \| v_{k-t-1} \|_{BL} E_{z,\gamma}^N [L(Z_t)] \alpha \bar{c} \right). \]

5.7.3 Proof of Step 2 in the Proof of Lemma 5.7.1

The claim is that

\[ E_{z,\gamma}^N [L(Z_t)] \leq E_{z,\gamma} [L(Z_t)] + 3 \alpha \bar{c} \sum_{m=0}^{t-1} (4 \alpha \bar{c} + 1)^m E_{z,\gamma} [L(Z_{t-m-1})]. \]

We first write

\[ E_{z,\gamma}^N [L(Z_t)] \leq E_{z,\gamma} [L(Z_t)] + \left( \| L \|_{BL} \right) \rho_{BL} \left( P_{z,t}^{\gamma}, P_{z,t}^{\gamma,N} \right) \]

(5.26)
where

\[ P_{z,t}^\gamma, P_{z,t}^\gamma,N \]

are the marginal distributions of the state \( z_t \) for the true model and approximate model respectively, with \( Z_0 = z \).

We focus on the term \( \| L \|_{BL} \rho_{BL} \left( P_{z,t}^\gamma, P_{z,t}^\gamma,N \right) \). We first claim that \( \| L \|_{BL} \leq 3 \), where \( L(z) = \rho(z, F(z)) \). Recall that \( F \) determines quantization given in (5.4).

\[
\| L \|_{BL} \leq \| L \|_{\infty} + \sup_{z,z'} |\rho(z, F(z)) - \rho(z', F(z'))| \leq 2 + \sup_{z,z'} |\rho(z, F(z)) - \rho(z', F(z'))| \rho(z, z')
\]

where we used the fact that

\[
\| L \|_{\infty} = \sup_z \rho_{BL}(z, F(z)) \leq 2
\]

as for any two probability measures \( \mu, \nu \), we have that \( \rho_{BL}(\mu, \nu) \leq 2 \) (see (4.6)).

Note that \( F \) is a nearest neighbor quantizer as defined in (5.4), thus we can write that

\[
|\rho(z, F(z)) - \rho(z', F(z'))| \leq \max (\rho(z, F(z')) - \rho(z', F(z')), \rho(z', F(z)) - \rho(z, F(z)))
\]

\[
\leq \sup_{\hat{z}} |\rho(z, \hat{z}) - \rho(z', \hat{z})| \leq \rho(z, z')
\]

where for the last step we used the triangle inequality. Hence we can conclude that

\[
\| L \|_{BL} \leq 2 + \sup_{z,z'} \frac{|\rho(z, F(z)) - \rho(z', F(z'))|}{\rho(z, z')} \leq 3.
\]
5.7. PROOFS OF TECHNICAL RESULTS

Now, we show that

\[
\rho_{BL} \left( P_{z,t}^\gamma, P_{z,t}^\gamma N \right) \leq \alpha_z \sum_{m=0}^{t-1} (4\alpha_z + 1)^m E_{z,\gamma} [L(Z_{t-m-1})]. \tag{5.28}
\]

We prove the claim by induction: for \( t = 1 \)

\[
\rho_{BL} \left( P_{z,1}^\gamma, P_{z,1}^\gamma N \right) = \sup_{\|f\|_{BL} \leq 1} \left| \int f(z_1) \eta(dz_1|F(z), u_1) - \int f(z_1) \eta(z_1) \eta(dz_1|z, u_1) \right| \\
\leq \alpha_z L(z).
\]

Now assume the claim holds for \( t - 1 \)

\[
\rho_{BL} \left( P_{z,t}^\gamma, P_{z,t}^\gamma N \right) \leq (1 + \alpha) \rho_{BL} \left( P_{z,t-1}^\gamma, P_{z,t-1}^\gamma N \right) + \alpha_z E_z^N [L(Z_{t-1})].
\]

Using (5.26), we can write

\[
\rho_{BL} \left( P_{z,t}^\gamma, P_{z,t}^\gamma N \right) \leq (1 + \alpha) \rho_{BL} \left( P_{z,t-1}^\gamma, P_{z,t-1}^\gamma N \right) + \alpha_z \left( E_z [L(Z_{t-1})] + 3 \rho_{BL} \left( P_{z,t-1}^\gamma, P_{z,t-1}^\gamma N \right) \right) \\
= (4\alpha_z + 1) \rho_{BL} \left( P_{z,t-1}^\gamma, P_{z,t-1}^\gamma N \right) + \alpha_z E_z [L(Z_{t-1})].
\]
\[ \leq (4\alpha_z + 1) \left( \alpha_z \sum_{m=0}^{t-2} (4\alpha_z + 1)^m E_z[L(Z_{t-m-2})] \right) + \alpha_z E_z[L(Z_{t-1})] \]

\[ = \alpha_z \sum_{m=0}^{t-1} (4\alpha_z + 1)^m E_z[L(Z_{t-m-1})] \]

which completes the proof of (5.28).

Now, we go back to the main claim and start from (5.26) to write

\[ E_{z,\gamma}^N [L(Z_t)] \leq E_{z,\gamma} [L(Z_t)] + \|L\|_{BL} \rho_{BL} \left( P_{z,t}^\gamma, P_{z,t}^{\gamma,N} \right) \leq E_{z,\gamma} [L(Z_t)] + 3\rho_{BL} \left( P_{z,t}^\gamma, P_{z,t}^{\gamma,N} \right) \]

\[ \leq E_{z,\gamma} [L(Z_t)] + 3\alpha_z \sum_{m=0}^{t-1} (4\alpha_z + 1)^m E_{z,\gamma} [L(Z_{t-m-1})]. \]

5.7.4 **Proof of Lemma 5.7.5**

**Lemma 5.7.4.** We introduce the following notation

\[ \hat{E}_z[L(Z_t)] = \sup_{u_0} \ldots \sup_{u_{t-2}} \int L(z_t) \eta(dz_t|z_{t-1}, u_{t-1}) \eta(dz_{t-1}|z_{t-2}, u_{t-2}) \ldots \eta(dz_1|z, u_0). \]

Under Assumption 5.3.1 we have that

\[ \sup_{\gamma \in \Gamma} E_{z,\gamma}[L(Z_t)] = \hat{E}_z[L(Z_t)]. \]

**Proof.** Recall that for a fixed policy \( \gamma = \{\gamma_t\}_t, \)

\[ E_{z,\gamma}[L(Z_t)] = \int L(z_t) \eta(dz_t|z_{t-1}, \gamma_{t-1}(z_{t-1})) \eta(dz_{t-1}|z_{t-2}, \gamma_{t-2}(z_{t-2})) \ldots \eta(dz_1|z, \gamma_0(z_0)) \]
It is easy to see that
\[
\int L(z_t) \eta(dz_t|z_{t-1}, \gamma_{t-1}(z_{t-1})) \eta(dz_{t-1}|z_{t-1}, \gamma_{t-2}(z_{t-2})) \ldots \eta(dz_1|z, \gamma_{0}(z_0))
\]
\[
\leq \int \sup_{u_{t-1}} \int L(z_t) \eta(dz_t|z_{t-1}, u_{t-1}) \eta(dz_{t-1}|z_{t-1}, \gamma_{t-2}(z_{t-2})) \ldots \eta(dz_1|z, \gamma_{0}(z_0))
\]
By repeating this step we can have
\[
\int L(z_t) \eta(dz_t|z_{t-1}, \gamma_{t-1}(z_{t-1})) \eta(dz_{t-1}|z_{t-1}, \gamma_{t-2}(z_{t-2})) \ldots \eta(dz_1|z, \gamma_{0}(z_0))
\]
\[
\leq \sup_{u_0} \int \ldots \sup_{u_{t-2}} \int \sup_{u_{t-1}} \int L(z_t) \eta(dz_t|z_{t-1}, u_{t-1}) \eta(dz_{t-1}|z_{t-2}, u_{t-2}) \ldots \eta(dz_1|z, u_0)
\]
Hence by taking supremum over all policies we can write
\[
\sup_{\gamma \in \Gamma} E \left[ L(Z_t) \right] \leq \hat{E}_z[L(Z_t)].
\]
For the other direction, we first focus on the term \( \sup_{u_{t-1}} \int L(z_t) \eta(dz_t|z_{t-1}, u_{t-1}) \), using Assumption 5.3.1, one can show that there exists a measurable map \( \gamma_{t-1} \) such that
\[
\sup_{u_{t-1}} \int L(z_t) \eta(dz_t|z_{t-1}, u_{t-1}) = \int L(z_t) \eta(dz_t|z_{t-1}, \gamma_{t-1}(z_{t-1})).
\]
Using the same argument, we can see that there exists a sequence of measurable functions \( \{ \gamma_0, \gamma_1, \ldots, \gamma_{t-1} \} \) such that
\[
\sup_{u_0} \int \ldots \sup_{u_{t-2}} \int \sup_{u_{t-1}} \int L(z_t) \eta(dz_t|z_{t-1}, u_{t-1}) \eta(dz_{t-1}|z_{t-2}, u_{t-2}) \ldots \eta(dz_1|z, u_0)
\]
\[
= \int L(z_t) \eta(dz_t|z_{t-1}, \gamma_{t-1}(z_{t-1})) \eta(dz_{t-1}|z_{t-1}, \gamma_{t-2}(z_{t-2})) \ldots \eta(dz_1|z, \gamma_{0}(z_0)).
\]
Hence we can write that

\[ \hat{E}_z[L(Z_t)] \leq \sup_{\gamma \in \Gamma} E_{z,\gamma}[L(Z_t)] \]

which proves the main claim.

**Lemma 5.7.5. Under Assumption 5.3.1**

\[ \|g\|_{BL} \leq K_0(\beta, \alpha_z, \|\hat{c}\|_{\infty}) \frac{3}{1 - \alpha_z}. \]

where

\[ K_0(\beta, \alpha_z, \|\hat{c}\|_{\infty}) = (\alpha_c + \beta \alpha_z \|J^*_{\beta}\|_{BL}) \frac{1}{1 - \beta(4\alpha_z + 1)}. \]

**Proof.** Using Lemma 5.7.4, we can write that

\[
g(z) = \lim_{k \to \infty} \sup_{\gamma} \left( \alpha_c \sum_{t=0}^{k-1} \beta^t \left( E_{z,\gamma}[L(Z_t)] + 3\alpha_z \sum_{m=0}^{t-1} (4\alpha_z + 1)^m E_{z,\gamma}[L(Z_{t-m-1})] \right) 
+ \alpha_z \sum_{t=0}^{k-1} \beta^{t+1} \|v_{k-t-1}\|_{BL} \left( E_{z,\gamma}[L(Z_t)] + 3\alpha_z \sum_{m=0}^{t-1} (4\alpha_z + 1)^m E_{z,\gamma}[L(Z_{t-m-1})] \right) \right)
\]

\[
= \lim_{k \to \infty} \left( \alpha_c \sum_{t=0}^{k-1} \beta^t \left( \hat{E}_z[L(Z_t)] + 3\alpha_z \sum_{m=0}^{t-1} (4\alpha_z + 1)^m \hat{E}_z[L(Z_{t-m-1})] \right) 
+ \alpha_z \sum_{t=0}^{k-1} \beta^{t+1} \|v_{k-t-1}\|_{BL} \left( \hat{E}_z[L(Z_t)] + 3\alpha_z \sum_{m=0}^{t-1} (4\alpha_z + 1)^m \hat{E}_z[L(Z_{t-m-1})] \right) \right).
\]

Using the fact that \( \|L\|_{\infty} \leq 2 \) we can write

\[ \|g\|_{\infty} \leq \lim_{k \to \infty} \left( \alpha_c \sum_{t=0}^{k-1} \beta^t \left( 2 + 3\alpha_z \sum_{m=0}^{t-1} 2(4\alpha_z + 1)^m \right) \right) \]
\[ + \alpha z \sum_{t=0}^{k-1} \beta^{t+1} \|v_{k-t-1}\|_{BL} \left( 2 + 3\alpha z \sum_{m=0}^{t-1} 2(4\alpha z + 1)^m \right) \]
\[ \leq 2K_0(\beta, \alpha z, \alpha \tilde{c}, \|\tilde{c}\|_{\infty}). \]

where
\[ K_0(\beta, \alpha z, \alpha \tilde{c}, \|\tilde{c}\|_{\infty}) = (\alpha \tilde{c} + \beta \alpha z \|J_{\beta}^*\|_{BL}) \frac{1}{1 - \beta(4\alpha z + 1)}. \]

Next, we show that when defined as a function of \( z \), \( \|\hat{E}_z[L(Z_t)]\|_{BL} \leq 3^{1-\alpha z^{t+1}} \frac{1-\alpha z^{t}}{1-\alpha z} \). We follow an inductive approach. For \( t = 1 \):
\[ |\hat{E}_z[L(Z_1)] - \hat{E}_z[L(Z_1)]| \leq \sup_u |\int L(z_1)\eta(dz_1|z,u) - \int L(z_1)\eta(dz_1|z,\hat{u},u)| \]
\[ \leq \|L\|_{BL} \alpha z \rho_{BL}(z, \hat{z}) \leq 3\alpha z \rho_{BL}(z, \hat{z}) \]

where \( \|L\|_{BL} \leq 3 \) follows from (5.27). Hence we have that \( \|\hat{E}_z[L(Z_1)]\|_{BL} \leq 3 + 3\alpha z \).

Now we assume the claim holds for \( t - 1 \) and focus on the case for \( t \):
\[ |\hat{E}_z[L(Z_t)] - \hat{E}_z[L(Z_t)]| \]
\[ = \sup_{u_0} \int \ldots \int L(z_t)\eta(dz_t|z_{t-1},u_{t-1})\eta(dz_{t-1}|z_{t-2},u_{t-2}) \ldots \eta(dz_1|z,u_0) \]
\[ - \sup_{u_0} \int \ldots \int L(z_t)\eta(dz_t|z_{t-1},u_{t-1})\eta(dz_{t-1}|z_{t-2},u_{t-2}) \ldots \eta(dz_1|\hat{z},u_0) \]
\[ \leq \sup_{u_0} \left( \hat{E}_{z_1} [L(Z_t)] \eta(dz_1|z,u_0) - \hat{E}_{z_1} [L(Z_t)] \eta(dz_1|\hat{z},u_0) \right) \]
\[ \leq \alpha z \rho_{BL}(z, \hat{z}) \|\hat{E}_z[L(Z_{t-1})]\|_{BL} \leq \alpha z \rho_{BL}(z, \hat{z})3 \frac{1-\alpha z^{t}}{1-\alpha z}. \]
We then have that

$$
\|\hat{E}_Z[L(Z_t)]\|_{BL} \leq \|L\|_\infty + 3\alpha_z \frac{1 - \alpha_z^t}{1 - \alpha_z} \leq 3 + 3\alpha_z \frac{1 - \alpha_z^t}{1 - \alpha_z} \leq \frac{3}{1 - \alpha_z}.
$$

Thus, we can write that

$$
\|\hat{E}_Z[L(Z_t)] - \hat{E}_\hat{z}[L(Z_t)]\| \leq 3\alpha_z \frac{1 - \alpha_z^t}{1 - \alpha_z} \rho_{BL}(z, \hat{z}) \leq \frac{3\alpha_z}{1 - \alpha_z} \rho_{BL}(z, \hat{z}).
$$

Hence for any $z, \hat{z}$

$$
|g(z) - g(\hat{z})| \leq \lim_{k \to \infty} \left( \alpha_c \sum_{t=0}^{k-1} \beta^t \left( \|\hat{E}_Z[L(Z_t)] - \hat{E}_\hat{z}[L(Z_t)]\| + 3\alpha_z \sum_{m=0}^{t-1} (4\alpha_z + 1)^m \left|\hat{E}_Z[L(Z_{t-m-1})] - \hat{E}_\hat{z}[L(Z_{t-m-1})]\right| \right)
+ \alpha_z \sum_{t=0}^{k-1} \beta^{t+1} \|v_{k-t-1}\|_{BL} \left( \|\hat{E}_Z[L(Z_t)] - \hat{E}_\hat{z}[L(Z_t)]\| \right)
+ 3\alpha_z \sum_{m=0}^{t-1} (4\alpha_z + 1)^m \left|\hat{E}_Z[L(Z_{t-m-1})] - \hat{E}_\hat{z}[L(Z_{t-m-1})]\right| \right)
\leq \lim_{k \to \infty} \left( \alpha_c \sum_{t=0}^{k-1} \beta^t \left( \frac{3\alpha_z}{1 - \alpha_z} \rho_{BL}(z, \hat{z}) + 3\alpha_z \sum_{m=0}^{t-1} (4\alpha_z + 1)^m \frac{3\alpha_z}{1 - \alpha_z} \rho_{BL}(z, \hat{z}) \right)
+ \alpha_z \sum_{t=0}^{k-1} \beta^{t+1} \|v_{k-t-1}\|_{BL} \left( \frac{3\alpha_z}{1 - \alpha_z} \rho_{BL}(z, \hat{z}) + 3\alpha_z \sum_{m=0}^{t-1} (4\alpha_z + 1)^m \frac{3\alpha_z}{1 - \alpha_z} \rho_{BL}(z, \hat{z}) \right) \right)
\leq K_0(\beta, \alpha_z, \alpha_c, \|\hat{c}\|_\infty) \frac{3\alpha_z}{1 - \alpha_z} \rho_{BL}(z, \hat{z}).
$$

Hence, we have that (assuming $\alpha_z \leq 1$)

$$
\|g\|_{BL} \leq K_0(\beta, \alpha_z, \alpha_c, \|\hat{c}\|_\infty) (2 + \frac{3\alpha_z}{1 - \alpha_z}) \leq K_0(\beta, \alpha_z, \alpha_c, \|\hat{c}\|_\infty) \frac{3}{1 - \alpha_z}.
$$
5.7. PROOFS OF TECHNICAL RESULTS

5.7.5 Proof of Lemma 5.7.6

Lemma 5.7.6. Under Assumption 5.3.1 we have that

\[ \|J_{\beta}^*\|_{BL} \leq \frac{1}{1 - \beta \alpha_z} \left( \frac{\|\tilde{c}\|_{\infty}}{1 - \beta} + \alpha \right). \]

Proof. For any \( x, y \in Z \)

\[ |J_{\beta}^*(x) - J_{\beta}^*(y)| \leq \sup_u \left( |\tilde{c}(x, u) - \tilde{c}(y, u)| + \beta \left| \int J_{\beta}^*(z) \eta(dz|x, u) - \int J_{\beta}^*(z) \eta(dz|y, u) \right| \right) \]

\[ \leq \alpha \rho_{BL}(x, y) + \beta \|J_{\beta}^*\|_{BL} \alpha \rho_{BL}(x, y). \]

Thus, we can write that

\[ \|J_{\beta}^*\|_{BL} \leq \|J_{\beta}^*\|_{\infty} + \alpha \rho_{BL}(x, y) + \beta \alpha \|J_{\beta}^*\|_{BL} \leq \frac{\|\tilde{c}\|_{\infty}}{1 - \beta} + \alpha \rho_{BL}(x, y) + \beta \alpha \|J_{\beta}^*\|_{BL}. \]

Hence, rearranging the terms, we can write (provided \( \beta \alpha_z < 1 \))

\[ \|J_{\beta}^*\|_{BL} \leq \frac{1}{1 - \beta \alpha_z} \left( \frac{\|\tilde{c}\|_{\infty}}{1 - \beta} + \alpha \right). \]

\[ \square \]
5.7. PROOFS OF TECHNICAL RESULTS

5.7.6 Proof of Theorem 5.4.1

Lemma 5.7.7. We have that for any \( \pi, \hat{\pi} \in \mathcal{P}(X) \) and for any \((y, u)[t, t-N] := \{ y_t, \ldots, y_{t-N}, u_t, \ldots, u_{t-N} \}\)

\[ Y^N \times U^N \]

\[
\| P^\pi(Y_{t+1} \in \cdot | (y, u)[t, t-N]) - P^\hat{\pi}(Y_{t+1} \in \cdot | (y, u)[t, t-N]) \|_{TV} \\
\leq \| P^\pi(X_t \in \cdot | y_{[t,N]}, u_{[t-1,N]}) - P^\hat{\pi}(X_t \in \cdot | y_{[t,N]}, u_{[t-1,N]}) \|_{TV}
\]

Proof. Let \( f \) be a measurable function of \( Y \) such that \( \| f \|_\infty \leq 1 \). We write

\[
\int f(y_{t+1}) P^\pi(dy_{t+1}|(y, u)[t, t-N]) - \int f(y_{t+1}) P^\hat{\pi}(dy_{t+1}|(y, u)[t, t-N]) \\
= \int f(y_{t+1}) Q(dy_{t+1}|x_{t+1}) T(dx_{t+1}|x_t, u_t) P^\pi(dx_t|y_{[t,N]}, u_{[t-1,N]}) \\
- \int f(y_{t+1}) Q(dy_{t+1}|x_{t+1}) T(dx_{t+1}|x_t, u_t) P^\hat{\pi}(dx_t|y_{[t,N]}, u_{[t-1,N]}) \\
\leq \| P^\pi(X_t \in \cdot | y_{[t,N]}, u_{[t-1,N]}) - P^\hat{\pi}(X_t \in \cdot | y_{[t,N]}, u_{[t-1,N]}) \|_{TV}
\]

at the last step, we used the fact that \( \int f(y_{t+1}) Q(dy_{t+1}|x_{t+1}) T(dx_{t+1}|x_t, u_t) \) is bounded by 1 as a function of \( x_t \) for \( \| f \|_\infty \leq 1 \). Taking the supremum over all \( \| f \|_\infty \leq 1 \) concludes the proof. \( \square \)

Proof. Proof of the main theorem: We prove the result for \( N = 1 \), the general case follows from identical steps. Let \( \hat{z}_0 = (\pi_0^-, y_1, y_0, u_0) \). Then we write

\[
\tilde{J}^N_{\beta}(\hat{z}_0) = J^N_{\beta}(\hat{\pi}, y_1, y_0, u_0) \\
= \min_{u_1 \in U} \left( c(\hat{\pi}, y_1, y_0, u_0, u_1) + \beta \sum_{y_2 \in Y} J^N_{\beta}(\hat{\pi}, y_2, y_1, u_1) P^\hat{\pi}(y_2|y_1, y_0, u_1, u_0) \right).
\]
Furthermore,

\[ J^*_\beta(\hat{z}_0) = J^*_\beta(\pi^*_0, y_1, y_0, u_0) \]

\[ = \min_{u_1 \in \mathcal{U}} \left( \tilde{c}(\pi^*_0, y_1, y_0, u_0, u_1) + \beta \sum_{y_2 \in \mathcal{Y}} J^*_\beta(\pi^*_1(\pi^*_0, y_0, u_0), y_2, y_1, u_1) P_{\pi^*_0}(y_2|y_1, y_0, u_0, u_0) \right). \]

Note that, for any \( \pi \in \mathcal{P}(\mathcal{X}) \), we have

\[ \tilde{J}^N_{\beta}(\pi, y_2, y_1, u_1) = \tilde{J}^N_{\beta}(\hat{\pi}, y_2, y_1, u_1) = \tilde{J}^N_{\beta}(\hat{\pi}, y_2, y_1, u_1). \]

In particular, we have that

\[ J^N_{\beta}(\hat{\pi}, y_2, y_1, u_1) = \tilde{J}^N_{\beta}(\pi^*_1(\pi^*_0, y_0, u_0), y_2, y_1, u_1). \]

Hence, we can write the following

\[ |\tilde{J}^N_{\beta}(\hat{z}_0) - J^*_\beta(\hat{z}_0)| \leq \max_{u_1 \in \mathcal{U}} \left| \hat{c}(\hat{\pi}, y_1, y_0, u_0, u_1) - \hat{c}(\pi^*_0, y_1, y_0, u_0, u_1) \right| \]

\[ + \max_{u_1 \in \mathcal{U}} \beta \left| \sum_{y_2 \in \mathcal{Y}} J^N_{\beta}(\hat{\pi}, y_2, y_1, u_1) P_{\hat{\pi}}(y_2|y_1, y_0, u_0, u_0) - \sum_{y_2 \in \mathcal{Y}} J^N_{\beta}(\pi^*_0, y_2, y_1, u_1) P_{\pi^*_0}(y_2|y_1, y_0, u_0, u_0) \right| \]

\[ + \max_{u_1 \in \mathcal{U}} \beta \left| \sum_{y_2 \in \mathcal{Y}} \tilde{J}^N_{\beta}(\pi^*_1(\pi^*_0, y_0, u_0), y_2, y_1, u_1) - J^*_\beta(\pi^*_1(\pi^*_0, y_0, u_0), y_2, y_1, u_1) \right| P_{\pi^*_0}(y_2|y_1, y_0, u_0, u_0). \]

Note that, by the definition of \( \hat{c} \), we have

\[ |\hat{c}(\hat{\pi}, y_1, y_0, u_0, u_1) - \hat{c}(\pi^*_0, y_1, y_0, u_0, u_1)| \leq \|c\|_\infty \|P_{\hat{\pi}}(X_1 \in \cdot|y_1, y_0, u_0) - P_{\pi^*_0}(X_1 \in \cdot|y_1, y_0, u_0)\|_{TV}. \]
If we denote \( \hat{z}_1 = (\hat{\pi}^1(\hat{\pi}_0^-, y_0, u_0), y_2, y_1, u_1) \), using Lemma 5.7.7 we can write

\[
E_{\pi_0}^{\gamma} \left[ |\tilde{J}_\beta^N(\hat{z}_0) - J_\beta^N(\hat{z}_0)| \right] \leq \|c\|_{\infty} E_{\pi_0}^{\gamma} \left[ \|P_{\hat{\pi}}(X_1 \in \cdot | Y_1, Y_0, U_0) - P_{\pi_0}^{\hat{\pi}}(X_1 \in \cdot | Y_1, Y_0, U_0)\|_{TV} \right] \\
+ \max_{u_1 \in \mathcal{U}} \beta \|J_\beta^N \|_{\infty} E_{\pi_0}^{\gamma} \left[ \|P_{\pi_0}^\gamma (y_2 | Y_1, Y_0, U_1, U_0) - P_{\hat{\pi}}(y_2 | Y_1, Y_0, U_1, U_0)\|_{TV} \right] \\
+ \max_{u_1 \in \mathcal{U}} \beta E_{\pi_0}^{\gamma} \left[ \sum_{y_2 \in \mathcal{Y}} |\tilde{J}_\beta^N(\hat{z}_1) - J_\beta^N(\hat{z}_1)| P_{\pi_0}^\gamma (y_2 | Y_1, Y_0, U_1, U_0) \right] \\
\leq \left( \|c\|_{\infty} + \beta \|J_\beta^N \|_{\infty} \right) L_0 + \max_{u_1 \in \mathcal{U}} \beta E_{\pi_0}^{\gamma} \left[ \sum_{y_2 \in \mathcal{Y}} |\tilde{J}_\beta^N(\hat{z}_1) - J_\beta^N(\hat{z}_1)| P_{\pi_0}^\gamma (y_2 | Y_1, Y_0, U_1, U_0) \right] \\
\leq \left( \|c\|_{\infty} + \beta \|J_\beta^N \|_{\infty} \right) L_0 + \sup_{\hat{\gamma} \in \Gamma} \beta E_{\pi_0}^{\gamma} \left[ |\tilde{J}_\beta^N(\hat{z}_1) - J_\beta^N(\hat{z}_1)| \right] \\

where

\[
L_t := \sup_{\hat{\gamma} \in \Gamma} E_{\pi_0}^{\gamma} \left[ \|P_{\hat{\pi}}^\gamma (X_{t+N} \in \cdot | Y_{[t,t+N]}, U_{[t,t+N-1]}) - P_{\pi_0}^\gamma (X_{t+N} \in \cdot | Y_{[t,t+N]}, U_{[t,t+N-1]})\|_{TV} \right]
\]

Then, following the same steps for \( E_{\pi_0}^{\gamma} \left[ |\tilde{J}_\beta^N(\hat{z}_1) - J_\beta^N(\hat{z}_1)| \right] \) and repeating the procedure, one can see that

\[
E_{\pi_0}^{\gamma} \left[ |\tilde{J}_\beta^N(\hat{z}_0) - J_\beta^N(\hat{z}_0)| \right] \leq \left( \|c\|_{\infty} + \beta \|J_\beta^N \|_{\infty} \right) \sum_{t=0}^{\infty} \beta^t L_t
\]

Note that \( \|J_\beta^N \|_{\infty} \leq \frac{\|c\|_{\infty}}{1 - \beta} \). Hence we can conclude

\[
E_{\pi_0}^{\gamma} \left[ |\tilde{J}_\beta^N(\hat{z}_0) - J_\beta^N(\hat{z}_0)| \right] \leq \frac{\|c\|_{\infty}}{(1 - \beta)} \sum_{t=0}^{\infty} \beta^t L_t.
\]
5.7. PROOFS OF TECHNICAL RESULTS

5.7.7 Proof of Theorem 5.4.2

Proof. We let \( \hat{z}_0 = (\pi_0^-, y_1, y_0, u_0) \). We denote the minimum selector for the approximate MDP by

\[
u^N_1 := \tilde{\phi}^N(\pi^-, y_1, y_0, u_0) = \phi^N(\hat{\pi}, y_1, y_0, u_0)
\]

and write

\[
J^\beta_\hat{\pi}(\hat{z}_0, \tilde{\phi}^N) = J^\beta(\pi^-, y_1, y_0, u_0, \tilde{\phi}^N)
\]

\[
= \tilde{c}(\pi^-, y_1, y_0, u_0, u^N_1) + \beta \sum_{y_2 \in \mathbb{Y}} J^\beta(\pi^-_1(y_0, u_0), y_2, y_1, u^N_1, \phi^N) P^\pi_{\pi^-}(y_2|y_1, y_0, u^N_1, u_0).
\]

Furthermore, we write the optimality equation for \( \tilde{J}^N_\beta \) as follows

\[
\tilde{J}^N_\beta(\hat{z}_0) = \tilde{c}(\hat{\pi}, y_1, y_0, u_0, u^N_1) + \beta \sum_{y_2 \in \mathbb{Y}} \tilde{J}^N_\beta(\pi^-_1(y_0, u_0), y_2, y_1, u^N_1) P^\beta(y_2|y_1, y_0, u^N_1, u_0).
\]

Hence, denoting \( \hat{z}_1 := (\pi^-_1(y_0, u_0), y_2, y_1, u^N_1) \) and using Lemma 5.7.7, we can write that

\[
E^\gamma_{\pi^-_0} \left[ |J^\beta_\hat{\pi}(\hat{z}_0, \tilde{\phi}^N) - \tilde{J}^N_\beta(\hat{z}_0)| \right] \leq \sup_{\tilde{\gamma} \in \Gamma} E^\gamma_{\pi^-_0} \left[ |\tilde{c}(\pi^-_0, Y_1, Y_0, U_0, U_1) - \tilde{c}(\hat{\pi}, Y_1, Y_0, U_0, U_1)| \right]
\]

\[
+ \sup_{\tilde{\gamma} \in \Gamma} E^\gamma_{\pi^-_0} \left[ \beta \sum_{y_2 \in \mathbb{Y}} J^\beta_\hat{\pi}(\hat{z}_1, \phi^N) P^\pi_{\pi^-}(y_2|Y_1, Y_0, U_0, U_1) - \beta \sum_{y_2 \in \mathbb{Y}} \tilde{J}^N_\beta(\hat{z}_1) P^\beta(y_2|Y_1, Y_0, U_1, U_0) \right]
\]

\[
\leq \|c\|_{\infty} \sup_{\tilde{\gamma} \in \Gamma} E^\gamma_{\pi^-_0} \left[ \|P^\pi_{\pi^-}(X_1 \in \cdot|Y_1, Y_0, U_0) - P^\pi_{\pi^-}(X_1 \in \cdot|Y_1, Y_0, U_0)\|_{TV} \right]
\]

\[
+ \beta \|\tilde{J}^N_\beta\|_{\infty} \sup_{\tilde{\gamma} \in \Gamma} E^\gamma_{\pi^-_0} \left[ \|P^\pi_{\pi^-}(y_2|Y_1, Y_0, U_1, U_0) - P^\beta(y_2|Y_1, Y_0, U_1, U_0)\|_{TV} \right]
\]
5.7. PROOFS OF TECHNICAL RESULTS

\[ + \beta \sup_{\hat{\gamma} \in \hat{\Gamma}} E_{\pi_0}^\hat{\gamma} \left[ \left| J_\beta(\hat{\gamma}_1, \tilde{\phi}^N) - \tilde{J}_\beta^N(\hat{\gamma}_1) \right| \right] \]
\[ \leq \|c\|_\infty L_0 + \beta \|\tilde{J}_\beta^N\|_\infty L_0 + \beta \sup_{\hat{\gamma} \in \hat{\Gamma}} E_{\pi_0}^\hat{\gamma} \left[ \left| J_\beta(\hat{\gamma}_1, \tilde{\phi}^N) - \tilde{J}_\beta^N(\hat{\gamma}_1) \right| \right]. \]

Following the same steps for \( E_{\pi_0}^\hat{\gamma} \left[ \left| J_\beta(\hat{\gamma}_1, \tilde{\phi}^N) - \tilde{J}_\beta^N(\hat{\gamma}_1) \right| \] and repeating the same procedure, with \( \|\tilde{J}_\beta^N\|_\infty \leq \frac{\|c\|_\infty}{1 - \beta} \) one can conclude that

\[ E_{\pi_0}^\hat{\gamma} \left[ \left| J_\beta(\hat{\gamma}_0, \tilde{\phi}^N) - \tilde{J}_\beta^N(\hat{\gamma}_0) \right| \right] \leq \frac{\|c\|_\infty}{(1 - \beta)} \sum_{t=0}^{\infty} \beta^t \beta^t. \quad (5.29) \]

Now, we go back to the theorem statement to write

\[ E_{\pi_0}^\hat{\gamma} \left[ \left| J_\beta(\hat{\gamma}_0, \tilde{\phi}^N) - J_\beta^*(\hat{\gamma}_0) \right| \right] \leq E_{\pi_0}^\hat{\gamma} \left[ \left| J_\beta(\hat{\gamma}_0, \tilde{\phi}^N) - \tilde{J}_\beta^N(\hat{\gamma}_0) \right| \right] + E_{\pi_0}^\hat{\gamma} \left[ \left| \tilde{J}_\beta^N(\hat{\gamma}_0) - J_\beta^*(\hat{\gamma}_0) \right| \right] \]
\[ \leq \frac{2\|c\|_\infty}{(1 - \beta)} \sum_{t=0}^{\infty} \beta^t. \]

The last step follows from (5.29) and Theorem 5.4.1
Chapter 6

Convergence of Finite Memory based Learning and Near Optimality under Filter Stability

In this chapter, we focus on approximate Q learning algorithms for partially observed models and for fully observed MDPs with continuous state spaces via quantizing their state spaces.

Q learning is a strong tool for model free learning of fully observed Markovian models with finite state spaces. The algorithm is based on learning Q values for every state and control action pair in an iterative fashion, where these Q values, in a way, serve as the ‘quality’ of the control action variables for the given state variable. Hence, based on the learned Q values, one can choose a policy which is almost surely optimal for the control problem under infinite horizon discounted cost criterion.

For the traditional Q learning algorithms, the assumptions that the state is fully observed and Markovian, and that the state space is finite, are vital for convergence of the iterations. Hence, applying the algorithm for partially observed models, and for the fully observed models with continuous state spaces is not immediately conclusive. In this chapter, we provide an algorithm for POMDPs where the Q values are constructed using finite
window information variables. Furthermore, we show that this approach can be used for fully observed MDPs with continuous state spaces via quantizing the state space, where the quantization is viewed as a measurement channel for a POMDP model and a history of unit window size is considered.

Another motivation for our study is the following: often one deals with problems where not only the specification of an MDP is unknown, but whether the problem is an MDP in the first place may not be known. The simplest extension perhaps is that of a POMDP where one is tempted to view the measurements as the state, or finite window of measurement and actions as the state. A question, which has not been resolved fully, is whether a Q-learning algorithm for such a setup would indeed converge, and the next question is if it does, what it converges to. Our answer to the first question is positive under mild conditions; and the second question is, under filter stability conditions, that the convergence is to near optimality with an explicit error bound between the performance loss and the memory window size.

In this chapter, as we have observed in Chapter 5, we will see, perhaps not surprisingly, that filter stability is an essential ingredient for the learning algorithm to arrive at optimal or near optimal solutions. In other words, how fast the process forgets its initial prior distribution when updated with the information variables will be a key aspect for the performance of the approximate Q values determined using most recent information variables. Unlike fully observed systems, the system (belief-MDP) states cannot be visited infinitely often for POMDPs since there are uncountably many belief states and the measurements collected should somehow present approximate information on the belief states through conditions related to filter stability. We will make this intuition precise in the following sections. We
also note that in optimal control theory, it is a standard result that (time-invariant) output feedback control performs poorly compared with state-feedback and, in the absence of observability, this holds for all memory lengths.

For this section, our focus will be on the infinite horizon discounted cost criterion for POMDPs where the cost is defined as

\[ J^*_\beta(\mu, T) = \inf_{\gamma \in \Gamma} J_{\beta}(\mu, T, \gamma), \]

with \( \mu \in \mathcal{P}(X) \) being the prior distribution of \( x_0 \). We also note that we use \( O(y|x) \) for the channel instead of \( Q(y|x) \) since \( Q \) is reserved for Q values in this chapter.

Before presenting the related literature and the approximate algorithm, we first briefly explain the Q learning algorithm for finite fully observed MDPs.

### 6.1 Q Learning Algorithm for Fully Observed MDPs

The traditional Q learning algorithm is an iteration of the following form

\[
Q_{k+1}(x, u) = (1 - \alpha_k(x, u))Q_k(x, u) + \alpha_k(x, u) \left( c(x, u) + \beta \min_{v} Q_k(X_1, v) \right)
\]

which is constructed for every \((x, u) \in X \times U\) pair using the data observed from a running system, and every time the \((x, u)\) pair is visited, the iteration is updated. Above, \( X_1 \) denotes the consecutive state variable that comes after the pair \((x, u)\), and \( c(x, u) \) is the cost realization of the \((x, u)\) pair, hence, the algorithm does not require any knowledge of the system dynamics or the cost function.
The following theorem establishes the optimality of the learned policies using the Q learning algorithm.

**Theorem 6.1.1** ([56][103]). Assume the following holds

- \( \sum_k \alpha_k = \infty, \sum_k \alpha_k^2 < \infty. \)
- Every \((x,u)\) pair is visited infinitely often during exploration.

Then, under any policy, the algorithm (6.2) converges almost surely for every \((x,u) \in X \times U\) to some \(Q^*(x,u)\) which satisfies

\[
Q^*(x,u) = c(x,u) + \beta \sum_{x_1} T(x_1|x,u) \min_v Q^*(x,v).
\]

Furthermore, any policy \(\gamma^*\) that satisfies \(Q^*(x,\gamma^*(x)) = \min_u Q^*(x,u)\), is optimal under the discounted cost expected cost criterion, that is, \(J_{\beta}(T,\gamma^*) = J^*_{\beta}(T)\).

Notice that, the algorithm in (6.2) is not applicable for POMDPs. However, a natural, though optimistic, suggestion to attempt to learn POMDPs would be to ignore the partial observability and pretend the noisy observations reflect the true state perfectly. For example, for infinite horizon discounted cost problems, one can construct Q iterations as:

\[
Q_{k+1}(y_k,u_k) = (1 - \alpha_k(y_k,u_k))Q_k(y_k,u_k) + \alpha_k(y_k,u_k) \left( C_k(y_k,u_k) + \beta \min_v Q_k(Y_{k+1},v) \right)
\]

(6.3)

where \(y_k\) represents the observations and \(u_k\) represents the control actions. We can further improve this algorithm by using not only the most recent observation but a finite window of past observations and control actions since we can infer information on the true state from the past data. Two main problems with this approach are that (i) first, the \((Y_k,U_k)\)
process is not a controlled Markov process (as only \((X_k, U_k)\) is) and the cost realizations \(C_k(y_k, u_k)\) depend on the observation process in a random and a time-dependent fashion, and hence the convergence of this approach does not follow directly from usual techniques ([56, 103]) and (ii) second, even if the convergence is guaranteed, it is not immediate what the limit Q values are, and whether they are meaningful at all. In particular, it is not known what MDP model gives rise to the limit Q values.

In the following parts of the chapter, namely in Section 6.4 we will prove that (6.3) converges almost surely to some fixed point equation, and the limit fixed point equation matches with the approximate model constructed in Section 5.4 using the finite window information variables. Hence, the filter stability conditions provided in the previous chapter will lead to the near optimality of the learned policies under finite window Q iterations.

In what follows, we first summarize the related literature, and then we present the approximate model from Section 5.4 again for convenience.

### 6.2 Literature Review

**On approximation methods.** The problem of approximate optimality is significantly more challenging compared to the fully observed counterpart as we have discussed in Chapter 5. Most of the studies in the literature are algorithmic and computational contributions. These include [85] and [122] which develop computational algorithms, utilizing structural convexity/concavity properties of the value function under the discounted cost criterion. [109] provides an insightful algorithm which may be regarded as a quantization of the belief space; however, no rigorous convergence results are provided. References [99] and [84] also present quantization based algorithms for the belief state, where the state, measurement, and the action sets are finite.
For partially observed setups, [94], building on [93], introduces a rigorous approximation analysis (and explicit methods for quantization of probability measures) after establishing weak continuity conditions on the transition kernel defining the (belief-MDP) via the non-linear filter [41, 61], and shows that finite model approximations obtained through quantization are asymptotically optimal and the control policies obtained from the finite model can be applied to the actual system with asymptotically vanishing error as the number of quantization bins increases. Another rigorous set of studies is [119] and [120] where the authors provide an explicit quantization method for the set of probability measures containing the belief states, where the state space is parametrically representable under strong density regularity conditions. The quantization is done through the approximations as measured by the Kullback-Leibler divergence (relative entropy) between probability density functions. [101] presents a notion of approximate information variable and studies near optimality of policies that satisfies the approximate information state property.

We refer the reader to the survey papers [73, 111, 48] and the recent book [67] for further structural results as well as algorithmic and computational methods for approximating POMDPs. Notably, for POMDPs [67] presents structural results on optimal policies under monotonicity conditions of the value function in the belief variable.

**On learning for POMDPs.** Learning in POMDPs is challenging for the reasons discussed above: if one attempts to learn optimal policies through empirical observations, then the analysis and convergence properties become significantly harder to obtain as the observations progress in a non-Markovian fashion and the belief state is uncountable. [57] studies a learning algorithm for POMDPs with average cost criterion where a policy improvement method is proposed using random polices and the convergence of this method to local optima is given. [76] and [72] are studies that propose the same approach as we use
6.2. LITERATURE REVIEW

in this chapter, where they use a finite memory of history to construct learning algorithms. They provide extensive experimental results, however, both lack a rigorous convergence or approximation result.

[98] studied (6.3), that is the Q learning algorithm for POMDPs by ignoring the partial observability and constructing the algorithm using the most recent observation variable (where the state, action and measurements spaces were all assumed finite), and established convergence of this algorithm under mild conditions (notably that the hidden state process is uniquely ergodic under the exploration policy which is random and puts positive measure to all action variables). In our work, we will consider memory sizes of more than 0 for the information variables and a continuous state space, and thus the algorithm in [98] can be seen as a special case of our setup. Different from our work, however, [98] does not study what the limit of the iterations mean, and in particular whether the limit equation corresponds to some MDP model. In this chapter, we rigorously construct the approximate belief MDP (the model constructed in Section 5) that the limit equation satisfies which gives an operational and practical conclusion regarding the analysis of the algorithm. Furthermore, we use different window sizes which turns out to be crucial for the performance of the learned policy: using longer window sizes reveals the intimate connection between the approximate learning problem and the nonlinear controlled filter stability problem that we will study in detail. This ultimately leads to near optimality of the $N$-window variation of (6.3) with an explicit approximation and robustness error bound as a function of $N$ and a computable/boundable coefficient related to filter stability.

On finite-memory approximations. We end the literature review section by mentioning two particularly related studies on finite-memory control for POMDPs that were already discussed in Chapter 5.
Reference [112] is a particularly related work that studies approximation techniques for POMDPs using finite memory with finite state, action, measurements. The POMDP is reduced to a belief MDP and the worst and best case predictors prior to the $N$ most recent information variables are considered to build an approximate belief MDP. The original value function is bounded using these approximate belief MDPs that use only finite memory, where the finiteness of the state space is critically utilized. Furthermore, a loss bound is provided for a suboptimally constructed policy that only uses finite history, where the bound depends on a specific ergodicity coefficient (which requires restrictive sample path contraction properties). In this chapter, we will consider more general signal spaces and consider more relaxed filter stability requirements, and, in particular, establish explicit rates of convergence results. We also rigorously construct the finite belief MDP considering the approximate Q learning algorithm whereas [112] only focuses on the approximation aspect of POMDPs.

In [114], the authors study near optimality of finite window policies for average cost problems where the state, action and observation spaces are finite; under the condition that the liminf and limsup of the average cost are equal and independent of the initial state, the paper establishes the near-optimality of (non-stationary) finite memory policies. Here, a concavity argument building on Feinberg [38] (which becomes consequential by the equality assumption) and the finiteness of the state space is crucial. The paper shows that for any given $\epsilon > 0$, there exists an $\epsilon$-optimal finite window policy. However, the authors do not provide a performance bound related to the length of the window, and in fact the proof method builds on convex analysis.
6.3 Approximate Belief MDP Construction with Finite Memory Feedback Revisited

In this section, we review the approximate belief MDP model from Section 5.4. We will later prove that the finite window Q iterations converge to a limit value which satisfies the optimality equation of the finite belief MDP model constructed via finite memory feedback variables. Hence, for convenience, we present the approximation method in this chapter again and present the key results.

6.3.1 An Alternative Finite Window Belief-MDP Reduction

We construct a fully observed MDP reduction with the condition that the controller has observed at least $N$ information variables, using the predictor from $N$ stages earlier and the most recent $N$ information variables (that is, measurements and actions). This new construction allows us to highlight the most recent information variables and compress the information coming from the past history via the predictor as a probability measure valued variable. In what follows, we will sometimes consider the case with $N = 1$ for some of the proofs to make the presentation less complicated. The general case follows from identical arguments.

Consider the following state variable at time $t$:

$$\hat{z}_t = (\pi_{t-N}, I_t^N)$$  \hspace{1cm} (6.4)

where, for $N \geq 1$

$$\pi_{t-N} = Pr(X_{t-N} \in \cdot | y_{t-N-1}, \ldots, y_0, u_{t-N-1}, \ldots, u_0),$$

$$I_t^N = \{y_t, \ldots, y_{t-N}, u_{t-1}, \ldots, u_{t-N}\}$$
and $I_t^N = y_t$ for $N = 0$ with $\mu$ being the prior probability measure on $X_0$. The state space with this representation is $\hat{Z} = \mathcal{P}(X) \times \mathbb{Y}^{N+1} \times \mathbb{U}^N$ where we equip $\hat{Z}$ with the product topology where we consider the weak convergence topology on the $\mathcal{P}(X)$ coordinate and the usual (coordinate) topologies on $\mathbb{Y}^{N+1} \times \mathbb{U}^N$ coordinates.

This new state representation can be mapped to the belief state $z_t$, which is defined as

$$z_t := Pr\{X_t \in \cdot | Y_0, \ldots, Y_t, U_0, \ldots, U_{t-1}\} \in \mathcal{P}(X).$$

Consider the map $\psi : \hat{Z} \to \mathcal{P}(X)$, for some $\hat{z}_t = (\pi_{t-N}, I_t^N)$

$$\psi(\hat{z}_t) = \psi(\pi_{t-N}, I_t^N) = P^{\pi_{t-N}}(X_t \in \cdot | I_t^N) = P^{\pi_{t-N}}(X_t \in \cdot | y_t, \ldots, y_{t-N}, u_{t-N-1}, \ldots, u_{t-N-1})$$

$$= P^{\mu}(X_t \in \cdot | y_t, \ldots, y_0, u_{t-1}, \ldots, u_0) = z_t$$

such that the map $\psi$ acts as a Bayesian update of $\pi_{t-N}$ using $I_t^N$. Using this map, we can define the stage-wise cost function and the transition probabilities. Consider the new cost function $\hat{c} : \hat{Z} \times U \to \mathbb{R}$, using the cost function $\tilde{c}$ of the belief MDP (defined in (4.5)) such that

$$\hat{c}(\hat{z}_t, u_t) = \hat{c}(\pi_{t-N}, I_t^N, u_t) = \tilde{c}(\psi(\pi_{t-N}, I_t^N), u_t)$$

$$= \int_X c(x_t, u_t) P^{\pi_{t-N}}(dx_t | y_t, \ldots, y_{t-N}, u_{t-1}, \ldots, u_{t-N}). \quad (6.5)$$

Furthermore, we can define the transition probabilities as follows: for some $A \in \mathcal{B}(\hat{Z})$ such that

$$A = B \times \{\hat{y}_{t-N+1}, \hat{u}_t, \ldots, \hat{u}_{t-N+1}\}, \quad B \in \mathcal{B}(\mathcal{P}(X))$$
we write

\[
\Pr(\hat{z}_{t+1} \in A | \hat{z}_t, \ldots, \hat{z}_0, u_t, \ldots, u_0) \\
= \Pr(\pi_{t-N+1}^{-1} \in B, \hat{y}_{t+1}, \ldots, \hat{y}_{t-N+1}, \hat{u}_t, \ldots, \hat{u}_{t-N+1} | \pi_{t-N}^{-1}, \ldots, \pi_0^{-1}, y_t, \ldots, y_0, u_t, \ldots, u_0) \\
= \mathbb{1}\{(y_t, \ldots, y_{t-N+1}, u_t, \ldots, u_{t-N+1}) = (\hat{y}_t, \ldots, \hat{y}_{t-N+1}, \hat{u}_t, \ldots, \hat{u}_{t-N+1})\} \\
\times \mathbb{1}\{G(\pi_{t-N}^{-1}, y_{t-N}, u_{t-N}) \in B\} \Pr(\pi_{t-N}^{-1} | y_t, \ldots, y_0, u_t, \ldots, u_0) \\
= \Pr(\pi_{t-N+1}^{-1} \in B, \hat{y}_{t+1}, \ldots, \hat{y}_{t-N+1}, \hat{u}_t, \ldots, \hat{u}_{t-N+1} | \pi_{t-N}^{-1}, y_t, \ldots, y_{t-N}, u_t, \ldots, u_{t-N}) \\
= \Pr(\hat{z}_{t+1} \in A | \hat{z}_t, u_t) \\
=: \int_A \hat{\eta}(d\hat{z}_{t+1} | \hat{z}_t, u_t)
\]

where the map \(G\) is defined as

\[
G(\pi_{t-N}^{-1}, y_{t-N}, u_{t-N}) = G(P^\mu(X_{t-N} \in \cdot | y_{t-N-1}, \ldots, y_0, u_{t-N-1}, \ldots, u_0, y_{t-N}, u_{t-N}) \\
= P^\mu(X_{t-N+1} \in \cdot | y_{t-N}, \ldots, y_0, u_{t-N}, \ldots, u_0).
\]

Hence, \(\hat{\eta}\) defines a controlled transition model for the new states \(\hat{z}_{t+1} \in \hat{Z}\). Then, we have a properly fully observed MDP, with the cost function \(\hat{c}\), transition kernel \(\hat{\eta}\) and the state space \(\hat{Z}\).

Note that any policy \(\phi : \mathcal{P}(X) \to U\) defined for the belief MDP, can be extended to the newly defined finite window belief-MDP using the map \(\psi\), and defining \(\hat{\phi} := \phi \circ \psi\) such that

\[
\hat{\phi}(\hat{z}) = \phi(\psi(z)).
\]

Thus, if an optimal policy can be found for the belief MDP, say \(\phi^*\), the policy \(\hat{\phi}^* = \phi^* \circ \psi\)
is an optimal policy for the newly defined MDP.

We now write the discounted cost optimality equation for the newly constructed finite window belief MDP. Note that with the alternative approach the state \( \hat{z} \) can only be written, if we have at least \( N \) information variables. Therefore, given that the decision maker observed at least \( N \) information variables, we write the following fixed point equation

\[
J^*_\beta(\hat{z}) = \min_{u \in U} \left( \hat{c}(\hat{z}, u) + \beta \int J^*_\beta(d\hat{z}_1|\hat{z}, u) \right).
\]

We can rewrite this fixed point equation in a different form, for notational ease assume \( N = 1 \). If \( \hat{z} \) has the form \( (\pi_0, y_1, y_0, u_0) \), then we can rewrite

\[
J^*_\beta(\pi_0, y_1, y_0, u_0) = \min_{u_1 \in U} \left( \hat{c}(\pi_0, y_1, y_0, u_0, u_1) + \beta \sum_{y_2 \in Y} J^*_\beta(\pi_1(\pi_0, y_0, u_0), y_2, y_1, u_1) P^{\pi_0}(y_2|y_1, y_0, u_1, u_0) \right).
\]

(6.6)

This representation will play an important role in the analysis of the problem. Note that the policy \( \hat{\phi}^* = \phi^* \circ \psi \) satisfies this fixed point equation.

The following fixed point equation can also be defined for any policy \( \hat{\phi} : \hat{Z} \to U \)

\[
J_\beta(\hat{z}, \hat{\phi}) = \hat{c}(\hat{z}, \hat{\phi}(\hat{z})) + \beta \int J_\beta(d\hat{z}_1|\hat{z}, \hat{\phi}(\hat{z}))
\]

where \( J_\beta(\hat{z}, \hat{\phi}) \) denotes the value function under the policy \( \hat{\phi} \) for the initial point \( \hat{z} \).
6.3. APPROXIMATE BELIEF MDP CONSTRUCTION WITH FINITE MEMORY
FEEDBACK REVISITED

6.3.2 Approximation of the Finite Window Belief-MDP

We now approximate the MDP constructed in the previous section. Consider the following set
\[ \hat{Z}_N^{\hat{\pi}} \]
for a fixed \( \hat{\pi} \in \mathcal{P}(X) \),
\[ \hat{Z}_N^{\hat{\pi}} = \{ (\hat{\pi}, y_{[0,N]}, u_{[0,N-1]}) : y_{[0,N]} \in Y^{N+1}, u_{[0,N-1]} \in U^N \} \] (6.7)
such that the state at time \( t \) is \( \hat{z}_t = (\hat{\pi}, I_N^t) \). Compared to the state \( \hat{z}_t = (\pi_t^{-N}, I_t^N) \) defined in (6.4), this approximate model uses \( \hat{\pi} \) as the predictor, no matter what the real predictor at time \( t - N \) is.

The cost function is defined in usual manner so that
\[ \hat{c}(\hat{z}_t^N, u_t) = \hat{c}(\hat{\pi}, I_t^N, u_t) = \hat{c}(\phi(\hat{\pi}, I_t^N), u_t) \]
\[ = \int_X c(x_t, u_t) P^{\hat{\pi}}(dx_t|y_t, y_{t-1}, u_t-1, u_t). \]

We define the controlled transition model by
\[ \hat{\eta}^N(\hat{z}_t+1_{t+1} | \hat{z}_t^N, u_t) = \hat{\eta}^N(\hat{\pi}, I_{t+1}^N | \hat{\pi}, I_t^N, u_t) := \hat{\eta}(\mathcal{P}(X), I_{t+1}^N | \hat{\pi}, I_t^N, u_t). \] (6.8)

For simplicity, if we assume \( N = 1 \), then the transitions can be rewritten for some \( I_{t+1}^N = (\hat{y}_{t+1}, \hat{y}_t, \hat{u}_t) \) and \( I_t^N = (y_t, y_t-1, u_t-1) \)
\[ \hat{\eta}^N(\hat{\pi}, \hat{y}_{t+1}, \hat{y}_t, \hat{u}_t | \hat{\pi}, y_t, y_{t-1}, u_t-1, u_t) = \hat{\eta}(\mathcal{P}(X), \hat{y}_{t+1}, \hat{y}_t, \hat{u}_t | \hat{\pi}, y_t, y_{t-1}, u_t-1, u_t) \]
\[ = 1_{\{y_t=\hat{y}_t, u_t=\hat{u}_t\}} \mathcal{P}^{\hat{\pi}}(\hat{y}_{t+1} | y_t, y_t-1, u_t, u_{t-1}). \] (6.9)

Denoting the optimal value function for the approximate model by \( J^N_\beta \), we can write
the following fixed point equation

\[ J^N_\beta(\hat{z}^N) = \min_{u \in U} \left( \hat{c}(\hat{z}^N, u) + \beta \sum_{\hat{z}_1^N \in \hat{Z}^N} J^N_\beta(\hat{z}_1^N) \hat{\eta}^N(\hat{z}_1^N | \hat{z}^N, u) \right). \quad (6.10) \]

By assuming \( N = 1 \) again, we can rewrite the fixed point equation for some \( \hat{z}_0^N = (\hat{\pi}, y_1, y_0, u_0) \) as

\[ J^N_\beta(\hat{\pi}, y_1, y_0, u_0) = \min_{u_1 \in U} \left( \hat{c}(\hat{\pi}, y_1, y_0, u_0, u_1) + \beta \sum_{y_2 \in Y} J^N_\beta(\hat{\pi}, y_2, y_1, u_1) P^\pi(y_2 | y_1, y_0, u_1, u_0) \right). \quad (6.11) \]

This representation will play a key role in the analysis of the problem investigated in this chapter. In particular, we will show that the finite window Q learning algorithm converges to this fixed point equation.

Since everything is finite in this setup, we can assume the existence of an optimal policy \( \phi^N \) that satisfies this fixed point equation. Note that both \( J^N_\beta \) and \( \phi^N \) are defined on the finite set \( \hat{Z}^N_\pi \). However, we can simply extend them to the set \( \hat{Z} \) by defining

\[ \tilde{J}^N_\beta(\hat{z}) = J^N_\beta(\pi, y_1, y_0, u_0) := J^N_\beta(\hat{\pi}, y_1, y_0, u_0) \]

\[ \tilde{\phi}^N(\hat{z}) = \phi^N(\pi, y_1, y_0, u_0) := \phi^N(\hat{\pi}, y_1, y_0, u_0) \]

for any \( \hat{z} = (\pi, y_1, y_0, u_0) \in \hat{Z} \).

Another point to note is that the policy \( \phi^N \) only uses most recent \( N \) information variables to choose the control actions.
Recall the following notation
\[
L_t := \sup_{\tilde{\gamma} \in \Gamma} \mathbb{E}^{\tilde{\gamma}}_{\pi_0} \left[ \left\| P^{\pi^\gamma} (X_{t+N} \in \cdot | Y_{[t,t+N]}, U_{[t,t+N-1]}) - P^\tilde{\gamma} (X_{t+N} \in \cdot | Y_{[t,t+N]}, U_{[t,t+N-1]}) \right\|_{TV} \right]
\]
(6.12)

which is the expected total variation distance between the posterior distributions of \(X_{t+N}\) conditioned on the same observation and control action variables \(Y_{[t,t+N]}, U_{[t,t+N-1]}\) when the prior distributions of \(X_t\) are given by \(\pi^{-t}\) and \(\hat{\pi}\).

**Theorem 6.3.1.** For \(\tilde{z}_0 = (\pi^{-0}, I_0^N)\), if a policy \(\hat{\gamma}\) acts on the first \(N\) step of the process which produces \(I_0^N\), we then have
\[
E^{\tilde{\gamma}}_{\pi_0} \left[ \left| J_\beta (\tilde{z}_0, \phi^N) - J_\hat{\beta} (\tilde{z}_0) \right| | I_0^N \right] \leq \frac{2 \|c\|_{\infty}}{(1 - \beta)} \sum_{t=0}^{\infty} \beta^t L_t.
\]

### 6.4 Q Iterations Using a Finite History of Information Variables and Convergence

Assume that we start keeping track of the last \(N+1\) observations and the last \(N\) control action variables after at least \(N+1\) time steps. That is, at time \(t\), we keep track of the information variables
\[
I_t^N = \begin{cases} 
\{y_t, y_{t-1}, \ldots, y_{t-N}, u_{t-1}, \ldots, u_{t-N}\} & \text{if } N > 0 \\
\{y_t\} & \text{if } N = 0.
\end{cases}
\]

We will construct the Q-value iteration using these information variables. In what follows, we will drop the \(N\) dependence on \(I_t^N\) and sometimes we will use \(N = 1\) for simplicity of the notation. For these new approximate states, we follow the usual Q learning algorithm
6.4. Q ITERATIONS USING A FINITE HISTORY OF INFORMATION VARIABLES AND CONVERGENCE

such that for any \( I \in Y^{N+1} \times U^N \) and \( u \in U \)

\[
Q_{k+1}(I, u) = (1 - \alpha_k(I, u))Q_k(I, u) + \alpha_k(I, u) \left( C_k(I, u) + \beta \min_v Q_k(I^k, v) \right),
\]

(6.13)

where \( I^k = \{Y_{t+1}, y_t, \ldots, y_{t-N+1}, u_t, \ldots, u_{t-N+1}\} \), we put the \( k \) dependence to emphasize that the distribution of \( Y_{t+1} \) and hence \( I^k \) are different for every \( k \), the time we hit \( \{y_t, y_{t-1}, \ldots, y_{t-N}, u_{t-1}, \ldots, u_{t-N}\} \) for the \( k \)-th time.

We note that the update times are different for each \((I, u)\) pair, that is, \( Q_k(I, u) \) is updated only when the process hits \((I, u)\).

To choose the control actions, we use policies that choose the control actions randomly and independent of everything else such that at time \( t \)

\[
u_t = u_i, \text{ w.p } \sigma_i
\]

for any \( u_i \in U \) with \( \sigma_i > 0 \) for all \( i \).

The algorithm differs from the usual Q-value iteration:

(i) The distribution of \( I^k_1 \), which is the consecutive N-window information variable when we hit the \((I, u)\) pair for the \( k \)-th time, is generally different for every \( k \) and the pair \((I, u)\) is not a controlled Markov process.

In other words, the controlled transitions are time dependent, that is, if we assume \( N = 1 \) and if we further assume that the \( k \)-th time we hit \((I, u)\) corresponds to the time step \( t \) in the real time flow of the process, then for some \( I = (y_t, y_{t-1}, u_{t-1}) \) and \( u = u_t \):

\[
Pr(I^k_1 = (y_{t+1}, y'_t, u'_t) | z = (y_t, y_{t-1}, u_{t-1}), u_t) = 1_{\{y_t = y'_t, u_t = u'_t\}} Pr(y_{t+1} | y_t, y_{t-1}, u_t, u_{t-1})
\]
is not stationary and might change at every time step \( t \), since \( \Pr(y_{t+1}|y_t, y_{t-1}, u_t, u_{t-1}) \) depends on the marginal distribution of \( x_{t-1} \) (\( x_{t-N} \) in the general case).

(ii) Here, we only observe the cost realizations of the underlying state process \( \{x_t\} \) and the control actions. For example, if we assume that \( N = 1 \) and that the \( k \)-th time we hit \( (I, u) \) corresponds to the time step \( t \) in the real time flow of the process, then the cost we observe is \( c(x_t, u_t) \). However, \( c(x_t, u_t) \) depends on \( (I, u) \) pair randomly and in a time dependent way so that for some \( I = (y_t, y_{t-1}, u_{t-1}) \) and \( u = u_t \):

\[
C_k(I, u) = c(x_t, u_t) \in B, \quad \text{w.p. } \Pr(X_t \in \{x : c(x, u_t) \in B\}|y_t, y_{t-1}, u_{t-1})
\]

where \( \Pr(dx_t|y_t, y_{t-1}, u_{t-1}) \) can be seen as some pseudo-belief on the underlying state variable given \( I = (y_t, y_{t-1}, u_{t-1}) \), the most recent \( N = 1 \) information variables. In other words, \( \Pr(dx_t|y_t, y_{t-1}, u_{t-1}) \) is the Bayesian update of \( \pi_{t-1} \), the marginal distribution of the true state \( x_{t-1} \) at the time step \( t-1 \), using \( I = (y_t, y_{t-1}, u_{t-1}) \) and thus, it is time dependent.

\[ \diamond \]

We will observe that, if one assumes that the hidden state process, \( \{x_t\} \) is positive Harris recurrent and in particular admits a unique invariant probability measure under stationary policies, say \( \hat{\pi} \) for the exploration policy \( \gamma \), then the average of approximate state transitions gets closer to

\[
P^*(I_{t+1}|I_t, u_t) := \hat{\eta}^N((\hat{\pi}, I_{t+1}))(\hat{\pi}, I_t, u_t)
\]  

(6.14)
6.4. Q ITERATIONS USING A FINITE HISTORY OF INFORMATION VARIABLES AND CONVERGENCE

with \( \hat{\eta}^N \) is defined as in (6.8) and (6.9). In particular, if we assume \( N = 1 \), then we write

\[
P^*(I_{t+1} = (y'_{t+1}, y'_t, u'_t)|I_t = (y_t, y_{t-1}, u_{t-1}, u_t)) = 1_{\{y'_t = y_t, u'_t = u_t\}} P^\hat{\pi}(y_{t+1}|y_t, y_{t-1}, u_t, u_{t-1})
\]

(6.15)

where \( P^\hat{\pi}(y_{t+1}|y_t, y_{t-1}, u_t, u_{t-1}) \) denotes the distribution of \( y_{t+1} \) when the marginal distribution on \( x_{t-1} \) is given by the invariant measure \( \hat{\pi} \).

We also have that the sample path averages of the random cost realizations get close to,

\[
C^*(I, u) = \hat{c}(\hat{\pi}, I, u) = \int_X c(x, u) P^\hat{\pi}(dx|I)
\]

where, \( P^\hat{\pi}(x|I) \) is the Bayesian update of \( \hat{\pi} \), using \( I \) and \( \hat{c}(\hat{\pi}, I, u) \) is defined as in (6.5). If we assume \( N = 1 \), we can write for some \( I = (y_1, y_0, u_0) \) and \( u = u_1 \)

\[
C^*(y_1, y_0, u_0, u_1) = \hat{c}(\hat{\pi}, (y_1, y_0, u_0), u_1) = \int_X c(x_1, u_1) P^\hat{\pi}(dx_1|y_1, y_0, u_0).
\]

(6.16)

Now consider the following fixed point equation

\[
Q^*(I, u) = C^*(I, u) + \beta \sum_{I'} P^*(I'|I, u) \min_v Q^*(I', v)
\]

(6.17)

where \( P^* \) is defined in (6.14) and \( C^* \) is defined in (6.16).

The existence of a such fixed point follows from usual contraction arguments. The same fixed equation can also be written as, for \( N = 1 \), and for \( I = (y_1, y_0, u_0) \) and \( u = u_1 \)

\[
Q^* ((y_1, y_0, u_0), u_1) = C^* ((y_1, y_0, u_0), u_1) + \beta \sum_{y_2 \in Y} P^\hat{\pi}(y_2|y_1, y_0, u_1, u_0) \min_{v \in U} Q^* ((y_2, y_1, u_1), v)
\]

(6.18)
For the rest of the chapter, we will use the following notation

\[ V^*(I) := \min_{v \in U} Q^*(I, v) \tag{6.19} \]

\[ V_t(I) := \min_{v \in U} Q_t(I, v). \tag{6.20} \]

We note that the stationary distribution \( \hat{\pi} \) does not have to be calculated by the decision maker. We will show that the algorithm naturally converges to (6.17), if the state process is positive Harris recurrent and in particular uniquely ergodic. The performance loss will depend on the stationary distribution \( \hat{\pi} \), however, we will establish further upper bounds that are uniform over such \( \hat{\pi} \) which decrease exponentially with the window size \( N \).

**Assumption 6.4.1.**

1. \( \alpha_t(I, u) = 0 \) unless \((I_t, u_t) = (I, u)\). Furthermore,

\[ \alpha_t(I, u) = \frac{1}{1 + \sum_{k=0}^{t-1} 1\{I_k = I, u_k = u\}} \]

We note that, this means that \( \alpha_k(I, u) = \frac{1}{k} \) if \( I_k = I, u_k = u \), if \( k \) is the instant of the \( k \)th visit to \((I, u)\), as this will be crucial in the averaging of the Markov chain dynamics (see Remark 6.4.1).

2. Under every stationary \{memoryless or finite memory exploration\} policy, say \( \gamma \), the true state process, \( \{X_t\}_t \), is positive Harris recurrent and in particular admits a unique invariant measure \( \pi^*_\gamma \).

3. During the exploration phase, every \((I, u)\) pair is visited infinitely often.

**Theorem 6.4.1.** Under Assumption 6.4.1
i. The algorithm given in (6.13) converges almost surely to $Q^*$ which satisfies (6.17).

ii. For any policy $\gamma^N$ that satisfies $Q^*(I,\gamma^N(I)) = \min_u Q^*(I,u)$, if we assume that the controller starts using $\gamma^N$ at time $t = N$ (after observing at least $N$ information variables), then denoting the prior distribution of $X_N$ by $\pi_0^N$, conditioned on the first $N$ step information variables we have

$$
E \left[ J_\beta(\pi_0^N, T, \gamma^N) - J_\beta^*(\pi_0^N, T) \right] \leq \frac{2\|c\|_\infty}{(1 - \beta)} \sum_{t=0}^{\infty} \beta^t L_t
$$

where $L_t$ is defined in (6.12) such that

$$
L_t := \sup_{\hat{\pi} \in \hat{\Gamma}} \mathbb{E}_{\pi_0} \left[ \| P^{\pi_0^t} (X_{t+N} \in \cdot | Y_{[t,t+N]}, U_{[t,t+N-1]}) - P^{\hat{\pi}^t} (X_{t+N} \in \cdot | Y_{[t,t+N]}, U_{[t,t+N-1]}) \|_{TV} \right]
$$

and $\hat{\pi}$ is the invariant measure on $x_t$ under the exploration policy $\gamma$.

Proof. For the proof of $i$, that is for the convergence of Q-learning, we separate the iterations into sub-iterations which are linear (as in [56], where this superposition principle of linear systems theory is utilized in showing the convergence of standard Q-learning algorithm). For the first part of the separated iterations, we use the fact that the dynamic programming equation is a contraction to prove its convergence which is similar to the traditional Q learning algorithms. For the remaining part of the iteration, we analyze the asymptotic behaviour of $I^k_1$, in which we distinguish our analysis from the traditional Q learning algorithms: For the usual Q iterations, one needs to study $X_1$ that is the consecutive state following some $(x,u)$ pair, and we have that $X_1 \sim \mathcal{T}(. | x,u)$. Thus, it is distributed independently and identically given $(x,u)$ which allows one to use Robbins–Monro type algorithms, to show the convergence. However, distributions of $I^k_1$'s are time dependent and not controlled Markovian. To study the asymptotic behavior of $I^k_1$, we construct a
different pair process which is Markov and we use ergodicity properties of Markov chains.

We first prove that the process $Q_k$, determined by the algorithm in (6.13), converges almost surely to $Q^*$. We define

$$
\Delta_k(I, u) := Q_k(I, u) - Q^*(I, u)
$$

$$
F_k(I, u) := C_k(I, u) + \beta V_k(I_k^1) - Q^*(I, u)
$$

$$
\hat{F}_k(I, u) := C^*(I, u) + \beta \sum_{I_1} V_k(I_1)P^*(I_1|I, u) - Q^*(I, u),
$$

where ($V_k$ is defined in [6.19]).

Then, we can write the following iteration

$$
\Delta_{k+1}(I, u) = (1 - \alpha_k(I, u))\Delta_k(I, u) + \alpha_k(I, u)F_k(I, u).
$$

Now, we write $\Delta_k = \delta_k + w_k$ such that

$$
\delta_{k+1}(I, u) = (1 - \alpha_k(I, u))\delta_k(I, u) + \alpha_k(I, u)\hat{F}_k(I, u)
$$

$$
w_{k+1}(I, u) = (1 - \alpha_k(I, u))w_k(I, u) + \alpha_k(I, u)r_k(I, u)
$$

where $r_k := F_k - \hat{F}_k = \beta V_k(I^k_1) - \beta \sum_{I_1} V_k(I_1)P^*(I_1|I, u) + C_k(I, u) - C^*(I, u)$. Next, we define

$$
r^*_k(I, u) = \beta V^*(I^k_1) - \beta \sum_{I_1} V^*(I_1)P^*(I_1|I, u) + C_k(I, u) - C^*(I, u)
$$
We further separate $w_k = u_k + v_k$ such that

$$u_{k+1}(I, u) = (1 - \alpha_k(I, u))u_k(I, u) + \alpha_k(I, u)e_k(I, u)$$

$$v_{k+1}(I, u) = (1 - \alpha_k(I, u))v_k(I, u) + \alpha_k(I, u)r^*_k(I, u)$$

where $e_k = r_k - r^*_k$.

In Section 6.8 we show that $v_k(I, u) \to 0$ almost surely for all $(I, u)$.

Now, we go back to the iterations:

$$\delta_{k+1}(I, u) = (1 - \alpha_k(I, u))\delta_k(I, u) + \alpha_k(I, u)\hat{F}_k(I, u)$$

$$u_{k+1}(I, u) = (1 - \alpha_k(I, u))u_k(I, u) + \alpha_k(I, u)e_k(I, u)$$

$$v_{k+1}(I, u) = (1 - \alpha_k(I, u))v_k(I, u) + \alpha_k(I, u)r^*_k(I, u).$$

Note that, we want to show $\Delta_k = \delta_k + u_k + v_k \to 0$ almost surely and we have that $v_k(I, u) \to 0$ almost surely for all $(I, u)$. The following analysis holds for any path that belongs to the probability one event in which $v_k(I, u) \to 0$. For any such path and for any given $\epsilon > 0$, we can find an $N < \infty$ such that $\|v_k\|_\infty < \epsilon$ for all $k > N$ as $(I, u)$ takes values from a finite set.

We now focus on the term $\delta_k + u_k$ for $k > N$:

$$(\delta_{k+1} + u_{k+1})(I, u) = (1 - \alpha_k(I, u))(\delta_k + u_k)(I, u) + \alpha_k(I, u)(\hat{F}_k + e_k)(I, u). \quad (6.21)$$

Observe that for $k > N$,

$$(\hat{F}_k + e_k)(I, u) = (F_k - r^*_k)(I, u) = \beta V_k(I^*_1) - \beta V^*(I^*_1) \leq \beta \max_{I, u} |Q_k(I, u) - Q^*(I, u)| = \beta \|\Delta_k\|_\infty$$
\[ \leq \beta \| \delta_k + u_k \|_\infty + \beta \epsilon \]

where the last step follows from the fact that \( v_k \to 0 \) almost surely. By choosing \( C < \infty \) such that \( \hat{\beta} := \beta(C + 1)/C < 1 \), for \( \| \delta_k + u_k \|_\infty > C \epsilon \), we can write that

\[ \beta \| \delta_k + u_k + \epsilon \|_\infty \leq \hat{\beta} \| \delta_k + u_k \|_\infty. \]

Now we rewrite (6.21)

\[
(\delta_{k+1} + u_{k+1})(I, u) = (1 - \alpha_k(I, u))(\delta_k + u_k)(I, u) + \alpha_k(I, u)(\hat{F}_k + e_k)(I, u) \\
\leq (1 - \alpha_k(I, u))(\delta_k + u_k)(I, u) + \alpha_k(I, u)\hat{\beta} \| \delta_k + u_k \|_\infty \\
< \| \delta_k + u_k \|_\infty
\]

Hence \( (\delta_{k+1} + u_{k+1})(I, u) \) clearly converges to 0 for \( \| \delta_k + u_k \|_\infty > C \epsilon \). This shows that the condition \( \| \delta_k + u_k \|_\infty > C \epsilon \) cannot be sustained indefinitely. Next, we show that once the process hits below \( C \epsilon \) it always stays there. Suppose \( \| \delta_k + u_k \|_\infty < C \epsilon \),

\[
(\delta_{k+1} + u_{k+1})(I, u) \leq (1 - \alpha_k(I, u))(\delta_k + u_k)(I, u) + \alpha_k(I, u)\beta \| \delta_k + u_k \|_\infty + \epsilon \\
\leq (1 - \alpha_k(I, u))C \epsilon + \alpha_k(I, u)\beta(C \epsilon + \epsilon) \\
= (1 - \alpha_k(I, u))C \epsilon + \alpha_k(I, u)\beta(C + 1) \epsilon \\
\leq (1 - \alpha_k(I, u))C \epsilon + \alpha_k(I, u)C \epsilon, \quad (\beta(C + 1) \leq C) \\
= C \epsilon.
\]

Then, we can write \( \| \delta_k + u_k \|_\infty < C \epsilon \). Hence, \( (\delta_k + u_k) \) process converges to some value below \( C \epsilon \) for any path that belongs to the probability one set.
Thus, taking $\epsilon \rightarrow 0$, we can conclude that $\Delta_k = \delta_k + u_k + v_k \rightarrow 0$ almost surely.

Therefore, the process $Q_k$, determined by the algorithm in (6.13), converges almost surely to $Q^*$.

For item (ii), recall that

$$Q^*(I, u) = C^*(I, u) + \beta \sum_{I_1} P^*(I_1 | I, u) \min_v Q^*(I, v).$$

This fixed point equation coincides with the DCOEs for the approximate belief MDP defined in (6.10) and (6.11). Hence, using Theorem 6.3.1, for any policy that satisfy $Q^*(I, \gamma^N(I)) = \min_u Q^*(I, u)$ we can write

$$E \left[ J_\beta(\pi^N_T, \gamma^N) - J_\beta^*(\pi^N_T, T) | I_0 \right] \leq \frac{2\|c\|_\infty}{(1 - \beta)} \sum_{t=0}^{\infty} \beta^t L_t$$

such that

$$L_t := \sup_{\gamma \in \Gamma} \left( \|P^* \pi_T^N (X_{t+N} \in \cdot | Y_{\left[t,t+N\right]}, U_{\left[t,t+N-1\right]}) - P^* \pi^N (X_{t+N} \in \cdot | Y_{\left[t,t+N\right]}, U_{\left[t,t+N-1\right]}) \|_{TV} \right)$$

and $\pi^*$ is the invariant measure on $x_t$ under the exploration policy $\gamma$.

**Remark 6.4.1.** The learning rates for the standard Q-learning algorithm require:

- $\sum_k \alpha_k = \infty$
- $\sum_k \alpha_k^2 < \infty$.

In our case, we have a particular form. To justify this, we note that although these standard two assumptions on the learning rates may be sufficient for convergence of the algorithm, the limit fixed point equation (if one exists) will not necessarily be useful. Consider the
following example where the state space is \( X = \{-1, +1\} \) and transitions are deterministic such that \( Pr(x_{t+1} = 1|x_t = -1) = 1, Pr(x_{t+1} = -1|x_t = +1) = 1 \) (leading to a periodic Markov chain). If one chooses the learning rates as \( \alpha_{2k} = 0, \alpha_{2k+1} = \sigma_k \) for every \( k \) such that \( \sigma_k \) is square summable but not summable, then even though the algorithm will converge, depending on the initial point, one of the transition models will always dominate the other. To avoid such examples, we choose the learning rates to be ”averaging” through time.

\[ \text{Remark 6.4.2.} \] We caution the reader that our result assumes that the cost starts running after time \( N \): that is the effective cost is:

\[
E \left[ \sum_{k=N}^{\infty} \beta^{k-N} c(x_k, u_k) \right].
\]  

(6.22)

Of course, this criterion is also applicable if the system starts running prior to time \( -N \) and the costs become in effect after time 0.

If this criterion is not applicable, and the first \( N \) stages are also crucial, (i) if \( \beta \) is large enough, we can conclude that the first \( N \) stages are not as critical for the analysis as their contributions will be minor in comparison with the future stages for the criterion, which can also be seen by considering this equivalent criterion to (6.1) and noting that for large enough \( \beta \), the contributions of the first \( N \) time stages become negligible:

\[
(1 - \beta) E \left[ \sum_{k=0}^{\infty} \beta^k c(x_k, u_k) \right].
\]

(ii) On the other hand, if \( \beta \) is not large and if the cost starts running at time 0, then, we can first run the \( Q \)-learning algorithm above to find the best \( N \)-window policies which
6.5. CONVERGENCE TO NEAR OPTIMALITY UNDER FILTER STABILITY

optimizes (6.22). The remaining question would be to optimize:

\[ E \left[ \sum_{k=0}^{N-1} c(x_k, u_k) + V(I_k) \right] \]

as a finite-horizon optimal control problem with a terminal cost and the terminal cost \( V \) can be estimated by (6.22) via Theorem 5.4.1 and Theorem 6.4.1. The question then becomes how to select the first \( N \) actions, leading to a problem with a finite search complexity for a finite horizon problem, without knowing the system dynamics. For this, one can run a MCMC algorithm in parallel simulations to find the optimal policy for the first \( N \) time stages. Since the resulting policy minimizing (6.23) will be at least as good as the first \( N \)-window policy under the optimal (belief-MDP) policy (which is not designed to optimize (6.23) but the original cost (6.1)), the bounds presented in Theorem 6.3.1 will be applicable even when the cost criterion includes the first \( N \) time stages.

6.5 Convergence to Near Optimality under Filter Stability

Recall that for Theorem 5.4.3, we used that

\[ L_t = \sup_{\tilde{\gamma} \in \tilde{\Gamma}} \mathbb{E}_{\pi_0}^{\tilde{\gamma}} \left[ \| P^{\pi_{\tilde{\gamma}}}(X_{t+N} \in \cdot | Y_{[t,t+N]}, U_{[t,t+N-1]}) - P^\pi(X_{t+N} \in \cdot | Y_{[t,t+N]}, U_{[t,t+N-1]}) \|_{TV} \right] \leq 2\alpha^N \]

for all \( t \), if \( \pi_{\tilde{\gamma}} \ll \tilde{\pi} \) where \( \alpha := (1 - \tilde{\delta}(T))(2 - \delta(Q)) \) (see Definition 5.2.3).

The absolute continuity assumption, that is \( \pi_{\tilde{\gamma}} \ll \tilde{\pi} \), can be interpreted as follows: assume that the true starting distribution is \( \pi_{\tilde{\gamma}} \) but we start the update from an incorrect prior \( \tilde{\pi} \). The error can be fixed with the information, \( y_{[0,t]}, u_{[0,t-1]} \) eventually, as long as,
the incorrect starting distribution $\tilde{\pi}$, puts on a positive measure to every event that the real
starting distribution $\pi_t^-$ puts on a positive measure. However, if it is not the case, that is,
if the incorrect starting distribution $\tilde{\pi}$ puts 0 measure to some event, that $\pi_t^-$, puts positive
measure to, information variables are not sufficient to fix the starting error occurring from
that 0 measure event. Of course this would not be feasible as the prior would not be com-
patible with the measured data. In any case, in our setup, the incorrect prior serves as an
approximation and this can be made to satisfy the absolute continuity condition by design:
this will be the invariant measure on the state under the exploration policy.

Recall that the Q learning iteration that uses finite window information variables, learns
the Q values for approximate states of the form $(\tilde{\pi}, I_t^N)$, instead of the true states $(\pi_{t-N}^-, I_t^N)$.
Inequality (6.24) suggests that the approximation error arising from using the stationary
distribution, $\tilde{\pi}$, instead of $\pi_{t-N}^-$, can be fixed with the information variables $I_t^N$, if $\tilde{\pi}$ cap-
tures the non zero events of $\pi_{t-N}^-$, that is if $\pi_{t-N}^- \ll \tilde{\pi}$.

Corollary 6.5.1 (to Theorem 6.4.1). Assume the following holds:

- Assumption 6.4.1 holds.
- The state space, $X$, is finite.
- Under the exploring policy, $\gamma$, the state process $\{x_t\}_t$ is irreducible.
- $\alpha := (1 - \tilde{\delta}(T))(2 - \delta(O)) < 1$.

Then, for any policy $\gamma^N$ that satisfies $Q^*(I, \gamma^N(I)) = \min_u Q^*(I, u)$, if we assume that
the controller starts using $\gamma^N$ at time $t = N$ (after observing at least $N$ information vari-
ables), then denoting the prior distribution of $X_N$ by $\pi_N^-$, conditioned on the first $N$ step
information variables we have

\[ E \left[ J_\beta(\pi_N, \mathcal{T}, \gamma^N) - J_\beta^*(\pi_N, \mathcal{T}) | I_0^N \right] \leq \frac{4\|c\|_\infty}{(1 - \beta)^2} \alpha^N. \]

Proof. Note that, by Theorem 6.4.1,

\[ E \left[ J_\beta(\pi_N, \mathcal{T}, \gamma^N) - J_\beta^*(\pi_N, \mathcal{T}) | I_0^N \right] \leq \frac{2\|c\|_\infty}{(1 - \beta)} \sum_{t=0}^{\infty} \beta^t L_t \]

If the state process \( x_t \) is irreducible under the exploring policy, then by Kac’s Lemma (59), we have that

\[ \hat{\pi}(x) > 0, \quad \forall x \in X. \]

Hence, using the inequality (6.24), we complete the proof.

**Corollary 6.5.2** (to Theorem 6.4.1). Assume the following holds:

- Assumption 6.4.1 holds.
- \( X \subset \mathbb{R}^m \) for some \( m < \infty \).
- The transition kernel \( \mathcal{T}(\cdot | x_0, u_0) \) admits a density function \( f \) with respect to a measure \( \phi \) such that \( \mathcal{T}(dx_1 | x_0, u_0) = f(x_1, x_0, u_0)\phi(dx_1) \) and \( f(x_1, x_0, u_0) > 0 \) for all \( x_1, x_0, u_0 \).
- \( \alpha := (1 - \bar{\delta}(\mathcal{T}))(2 - \delta(O)) < 1. \)

Then, for any policy \( \gamma^N \) that satisfies \( Q^*(I, \gamma^N(I)) = \min_u Q^*(I, u) \), if we assume that the controller starts using \( \gamma^N \) at time \( t = N \) (after observing at least \( N \) information variables), then denoting the prior distribution of \( X_N \) by \( \pi_N^{-} \), conditioned on the first \( N \) step
6.6. QUANTIZATION AS A MEASUREMENT CHANNEL: Q-LEARNING FOR CONTINUOUS SPACE MDPS VIA QUANTIZATION AND NEAR OPTIMALITY

In this section, we consider a fully observed system with a continuous state space and construct an approximate Q learning algorithm by discretizing the state space and using a finite subset of the state space for the Q iterations. We assume that the state space $X \subset \mathbb{R}^d$ is compact, and for a given $\epsilon > 0$, we quantize it by choosing a finite subset $\hat{X} = \{x_1, \ldots, x_n\}$ such that

$$\max_{x \in X} \min_{\hat{x} \in \hat{X}} |x - \hat{x}| \leq \alpha (1/n)^{1/d}$$

Proof. Note that, by assumption $T(dx|x_0, u_0) = f(x_1, x_0, u_0)\phi(dx_1)$ and $f(x_1, x_0, u_0) > 0$ for all $x_1, x_0, u_0$ and hence, under the exploration policy $\gamma$, the state process $x_t$ is $\phi$-irreducible and admits a unique invariant measure, say $\hat{\pi}$. Using the assumptions, we can also write that for any $A \in \mathcal{B}(X)$ with $\phi(A) > 0$

$$\hat{\pi}(A) = \int_Z \int_A \int_U f(x_1, x_0, u_0)\gamma(du_0)\phi(dx_1)\hat{\pi}(dx_0) > 0$$

which implies that $\phi \ll \hat{\pi}$. Note that the transition kernel $T(\cdot|x, u)$ is absolutely continuous with respect to $\phi$ for every $(x, u)$, and thus, for the predictor $\pi_t$ at any time step $t$, we can write that $\pi_t \ll \phi \ll \hat{\pi}$.

Hence, inequality (6.24) and Theorem 6.4.1 concludes the proof. \qed

6.6 Quantization as a Measurement Channel: Q-Learning for Continuous Space MDPs via Quantization and Near Optimality

In this section, we consider a fully observed system with a continuous state space and construct an approximate Q learning algorithm by discretizing the state space and using a finite subset of the state space for the Q iterations. We assume that the state space $X \subset \mathbb{R}^d$ is compact, and for a given $\epsilon > 0$, we quantize it by choosing a finite subset $\hat{X} = \{x_1, \ldots, x_n\}$ such that

$$\max_{x \in X} \min_{\hat{x} \in \hat{X}} |x - \hat{x}| \leq \alpha (1/n)^{1/d}$$
for some $\alpha > 0$, which is possible since $X$ is compact.

We use a nearest neighbor map $\rho : X \to \hat{X}$ to choose elements from the finite set $\hat{X}$ such that at any time instance $t < \infty$, if the state is $x_t$, we use

$$\hat{x}_t = \rho(x_t) := \arg \min_{\hat{x} \in \hat{X}} \|\hat{x} - x_t\|$$ (6.25)

for the Q learning algorithm. Note that with this map, we separate $X$ into $n$ subsets $\{B_1, \ldots, B_n\}$ such that for $x_i \in \hat{X}$, $B_i := \{x \in X : \rho(x) = x_i\}$.

### 6.6.1 Construction of a Finite MDP

We now present a finite MDP construction method that is taken from [91, Chapter 4] using the nearest neighbor map $\rho$.

Recall that the finite state space, we use is

$$\hat{X} = \{x_1, \ldots, x_n\}$$

where

$$\max_{x \in X} \min_{\hat{x} \in \hat{X}} |x - \hat{x}| \leq \alpha(1/n)^{1/d}. \quad (6.26)$$

We fix a probability measure $\hat{\pi} \in \mathcal{P}(\hat{X})$ such that $\hat{\pi}(B_i) > 0$ for all $i \in \{1, \ldots, n\}$, and we define $\hat{\pi}_i$ as the concentration of $\hat{\pi}$ on the set $B_i$ such that, for all $A \subset B_i$

$$\hat{\pi}_i(A) := \frac{\hat{\pi}(A)}{\hat{\pi}(B_i)}.$$ 

We now define the cost function for the set $\hat{X}$, denoted by $C^*$. For any $\hat{x}_i \in \hat{X}$, if the
quantization bin $x_i$ belongs to is $B_i$, we set
\[
C^*(x_i, u) := \int_{B_i} c(x, u) \tilde{\pi}_i(dx).
\]

Furthermore, we define the transition probabilities for the finite model as follows: for some $\hat{x}_i \in \hat{X}$ and $\hat{x}_j \in X$ where $\hat{x}_i \in B_i$ and $\hat{x}_j \in B_j$, we set
\[
P^*(\hat{x}_j | \hat{x}_i, u) := \int_{B_i} T(B_j | x, u) \tilde{\pi}_i(dx).
\]

We then have a finite MDP with state space $\hat{X}$, action space $U$, cost function $C^*$ and transition model $P^*$. If we find an optimal policy which satisfies the discounted cost optimality equation of this finite model, and denote the optimal policy by $\gamma_n$, then we can extend this policy to the continuous state space $X$ as follows: let us denote the extended policy by $\hat{\gamma}_n$, then
\[
\hat{\gamma}_n(x) := \gamma_n(\rho(x))
\]
where $\rho$ is the nearest neighbor map defined in (6.25).

We can then state the following near optimality result.

**Theorem 6.6.1.** [91] Theorem 4.38] Assume the following holds:

- The one-stage cost function $c(x, u)$ is Lipschitz continuous in $x$ with constant $\alpha_c$, for all $u \in U$, that is, $\sup_{x \neq x'} \frac{|c(x, u) - c(x', u)|}{\|x - x'\|} \leq \alpha_c$

- The transition kernel $T(\cdot | x_0, u_0)$ admits a density function $f$ with respect to a measure $\phi$ such that $T(dx_1 | x_0, u_0) = f(x_1, x_0, u_0) \phi(dx_1)$, and $f$ is Lipschitz continuous in $x_0$ with constant $\alpha_T$. 
• $X \subset \mathbb{R}^d$ is compact.

If we denote the policy constructed for the finite model by $\gamma_n$ and its extended version to $X$ by $\hat{\gamma}_n$, then we have

$$J_{\beta}(T; \hat{\gamma}_n) - J^*_\beta(T) \leq \frac{\tau(\beta, \alpha_T) \alpha e^{\frac{1}{1-\beta \alpha_T}} + \frac{2\alpha}{1-\beta^2} 2\alpha (1/n)^{1/d}}{1-\beta}$$

where $\tau(\beta, \alpha_T) = (2+\beta)\beta \alpha_T + \frac{\beta^2 + 4\beta + 2}{(1-\beta)^2}$ and $(1/n)^{1/d}$ arises from the quantization error (recall (6.26)).

6.6.2 Approximate Q Learning for Continuous State Spaces

We now construct an approximate Q learning algorithm for fully observed MDPs with continuous state spaces.

Using the map $\rho$ defined in (6.25), we write for any $(x, u) \in X \times U$

$$Q_{k+1}(\rho(x), u) = (1 - \alpha_k(\rho(x), u))Q_k(\rho(x), u) + \alpha_k(\rho(x), u) \left(C_k(\rho(x), u) + \beta \min_v Q_k(\rho(X_1), v)\right)$$

(6.29)

that is for any true value of the state, we use its representative state from the finite set $\hat{X}$. To choose the control actions, we again use polices that choose the control actions randomly and independent of everything else with positive probability for every action. We now argue that this approximate iteration can be seen as a special case of the POMDP iteration (6.13) by considering the discretization as a quantizer channel. If we consider the finite set $\hat{X}$ as the observation space and define the observation channel as a quantizer such that

$$O(\hat{x}_i|x) = 1_{x \in B_i},$$

(6.30)
then the algorithm in (6.29) is the same algorithm as in (6.13) with $N = 0$.

Thus, we can use the set $\hat{X}$ to construct the Q learning algorithm. Using the quantizer channel (6.30) and Theorem 6.4.1 (i) for $N = 0$, if the state process $x_t$ admits an invariant measure $\hat{\pi}$ under the exploration policy, the algorithm converges to

$$Q^*(\hat{x}, u) = C^*(\hat{x}, u) + \beta \sum_{\hat{x}_1} P^*(\hat{x}_1 | \hat{x}, u) \min_v Q^*(\hat{x}_1, v)$$

where, for $\hat{x} \in B$ and $\hat{x}_1 \in B_1$, if we define $\hat{\pi}^*(A) := \frac{\hat{\pi}(A)}{\hat{\pi}(B)}$ for all $A \subset B$ with $\hat{\pi}$ being the invariant measure, the cost and transitions are defined as

$$C^*(\hat{x}, u) := \int_{\mathcal{X}} c(x, u) P(dx | \hat{x}) = \int_B c(x, u) \hat{\pi}^*(dx)$$

$$P^*(\hat{x}_1 | \hat{x}, u) := \int_{\mathcal{X}} \mathcal{T}(B_1 | x, u) P(dx | \hat{x}) = \int_B \mathcal{T}(B_1 | x, u) \hat{\pi}^*(dx).$$

Note that this model is the same one we constructed in previous section (6.27) and (6.28). In the previous section, however, we have required that $\hat{\pi}(B_i) > 0$ for all $i$. Hence, the invariant measure of the state process $x_t$ should put positive measure to the quantization bins. To guarantee this, we assume that the transition kernel $\mathcal{T}(\cdot | x_0, u_0)$ admits a density function $f$ with respect to a measure $\phi$ such that $\mathcal{T}(dx_1 | x_0, u_0) = f(x_1, x_0, u_0)\phi(dx_1)$, $f(x_1, x_0, u_0) > 0$ for all $x_1, x_0, u_0$

Thus, we can state the following result, using Theorem 6.4.1 and [91, Theorem 4.38]:

**Theorem 6.6.2.** Under Assumption 6.4.1 if the transition kernel $\mathcal{T}(\cdot | x_0, u_0)$ admits a density function $f$ with respect to a measure $\phi$ such that $\mathcal{T}(dx_1 | x_0, u_0) = f(x_1, x_0, u_0)\phi(dx_1)$, $f(x_1, x_0, u_0) > 0$ for all $x_1, x_0, u_0$ and $f$ is Lipschitz continuous in $x_0$ with constant $\alpha_T$ and if $c(x, u)$ is Lipschitz continuous in $x$ with constant $\alpha_c$, then the Q learning algorithm
6.7. NUMERICAL STUDY

In (6.29) converges and for any policy $\gamma^n$ that satisfies $Q^*(\hat{x}, \gamma^n(\hat{x})) = \min_u Q^*(\hat{x}, u)$, we have

$$J_\beta(\mathcal{T}, \gamma^n) - J_\beta^*(\mathcal{T}) \leq \frac{\tau(\beta, \alpha_T) \alpha E}{1 - \beta} + \frac{2\alpha}{1 - \beta} 2\alpha (1/n)^{1/d}$$

where $\tau(\beta, \alpha_T) = (2 + \beta)\beta \alpha_T + \frac{\beta^2 + 1\beta + 2}{(1 - \beta)^2}$ and $\alpha (1/n)^{1/d}$ arises from the quantization error (recall (6.26)).

6.7 Numerical Study

In this section, we present the example presented in Section 5.6.

We recall that in this model, we have $X, Y, U = \{0, 1\}$ with

$$x_t = \begin{cases} 
  1 & \text{machine is working at time } t \\
  0 & \text{machine is not working at time } t
\end{cases} \quad u_t = \begin{cases} 
  1 & \text{machine is being repaired at time } t \\
  0 & \text{machine is not being repaired at time } t
\end{cases}$$

The one stage cost function is given by

$$c(x, u) = \begin{cases} 
  R + E & x = 0, u = 1 \\
  E & x = 0, u = 0 \\
  0 & x = 1, u = 0 \\
  R & x = 1, u = 1
\end{cases}$$

where $R$ is the cost of repair and $E$ is the cost incurred by a broken machine.

We study the example with discount factor $\beta = 0.8$. The transition probabilities are
chosen as follows:

\[ Pr(x_1 = 0|x_0 = 0, u_0 = 0) = 0.9, \quad Pr(x_1 = 0|x_0 = 0, u_0 = 1) = 0.6 \]
\[ Pr(x_1 = 0|x_0 = 1, u_0 = 0) = 0.4, \quad Pr(x_1 = 0|x_0 = 1, u_0 = 1) = 0.1 \]

For the measurement channel:

\[ Pr(y = 0|x = 0) = 0.7, \quad Pr(y = 1|x = 1) = 0.7. \]

For the cost function, we choose \( R = 3 \), and \( E = 1 \).

For the exploration, we use a random policy such that \( Pr(\gamma(x) = 0) = \frac{1}{2} \) and \( Pr(\gamma(x) = 1) = \frac{1}{2} \) for all \( x \). Under this policy, \( x_t \) admits a stationary policy \( \hat{\pi}(\cdot) = 0.42\delta_0(\cdot) + 0.58\delta_1(\cdot) \).

We have proved in Theorem 6.4.1 that the Q iteration given by (6.13) converges to the Q-values of the approximate belief-MDP defined in (6.10). Defining

\[ V_t(I) := \min_{v \in \mathcal{U}} Q_t(I,v), \]

in the next graphs, we plot \( \sup_I |V_t(I) - J_N^N(\hat{\pi}, I)| \) for \( N = 0, 1, 2 \): The convergence of the Q-values can be seen as:
This setup gives $\alpha = (1 - \tilde{\delta}(T)) \times (2 - \delta(O)) = 0.7$. The following graph shows the error $J_\beta(\mu, T, \gamma^N) - J_\beta^*(\mu, T)$, expected total variation filter stability term and $\alpha^N$ terms. We scale all of them to make them start from the same point to emphasize the decrease rates.
6.8 Technical Results

Proof of $v_k(I, u) \to 0$

We will show that $v_k(I, u) \to 0$ almost surely for all $(I, u)$. We prove the claim only for $N = 1$ case for simplicity and let $I = (y_1, y_0, u_0)$ and $u = u_1$ for some $(y_1, y_0, u_0, u_1) \in Y^2 \times U^2$. The proof for general $N$ follows from essentially same steps. We have

$$v_{k+1}(I, u) = (1 - \alpha_k(I, u))v_k(I, u) + \alpha_k(I, u)r_k^*(I, u)$$

Note that the $v_k(I, u)$ values are updated only when the process hits $(I, u)$ pair. Thus, we define the following stopping times

$$\tau(k + 1) = \{\min t > \tau(k) : Y_t = y_1, Y_{t-1} = y_0, U_{t-1} = u_0, U_t = u_1\}$$

where $\tau(0) = 0$. The stopping times $\tau(k)$ represents the time instants that the process hits $I = (y_1, y_0, u_0)$ and control $u_1$ is applied at that time. Note that, since we assume that the process hits each $(I, u)$ pair infinitely often, the stopping times are bounded almost surely.
When $\alpha_k(I, u) = \frac{1}{k}$ for every $(I, u)$ pair, the problem reduces to

$$v_{k+1}(I, u) = \frac{1}{k} \sum_{k'=0}^{k-1} r^{*}_{k'}(I, u).$$

Recall that

$$r^{*}_{k}(I, u) = \beta V^{*}(I^k_1) - \beta \sum_{I_1} V^{*}(I_1) P^{*}(I_1|I, u) + C_k(I, u) - C^{*}(I, u).$$

Hence, we will first analyze the asymptotic behavior of

$$I^k_1 := (Y_{\tau(k) + 1}, Y_{\tau(k)}, U_{\tau(k)})$$

that is the consecutive pair after we hit $I = (y_1, y_0, u_0)$ if we apply the control $u_1$. To analyze the asymptotic behavior of this term, we will make use of the Markov chain theory. Note that $I^k_1 = (Y_{\tau(k) + 1}, Y_{\tau(k)}, U_{\tau(k)})$ by itself is not a Markov chain. We define

$$\hat{X}_k := (X_{\tau(k) - 1}, U_{\tau(k) - 1})$$

which is the state variable whose measurement is $y_0$ when the process hits $I = (y_1, y_0, u_0)$ and the control action applied at that time. We will show that the pair $(I^k_1, \hat{X}_k)$ forms a Markov chain, under the exploring policy $\gamma$ which is assumed to be stationary and independent of the past $X_t$, $Y_t$ and $U_t$ processes. Then, we will use stationary distribution of this Markov chain to analyze the asymptotic behavior of $I^k_1$.

Consider the joint process $(Y_{t+1}, Y_t, U_t, X_{t-1}, U_{t-1})$, defined by the real time flow of the process, rather than sampled points at the stopping times. We prove that this process is
a Markov chain:

\[
Pr(Y_{t+1}, Y_t, U_t, X_{t-1}, U_{t-1} | y_t, y_{t-1}, \ldots, y_0, u_{t-1}, \ldots, u_0, x_{t-2}, \ldots, x_0) = \mathbb{1}_{\{Y_t = y_t\}} \mathbb{1}_{\{U_{t-1} = u_{t-1}\}} Pr(Y_{t+1} | y_t, x_{t-1}, u_t, u_{t-1}) Pr(X_{t-1} | y_{t-1}, x_{t-2}, u_{t-2}) \gamma(U_t)
\]

\[
= Pr(Y_{t+1}, Y_t, U_t, X_{t-1}, U_{t-1} | y_t, y_{t-1}, x_{t-2}, u_{t-1}, u_{t-2})
\]

where we use \(\gamma\) as the probability measure of the exploring policy which is independent of everything by assumption. Above, we used that

\[
Pr(Y_{t+1} | y_t, x_{t-1}, u_t, u_{t-1}) = Pr(Y_{t+1} | y_t, x_{t-2}, u_{t-2})
\]

\[
Pr(X_{t-1} | y_{t-1}, x_{t-2}, u_{t-1}, u_{t-2}) = Pr(X_{t-1} | y_{t-1}, x_{t-2}, u_{t-2})
\]

Since, the joint process \((Y_{t+1}, Y_t, U_t, X_{t-1}, U_{t-1})\) is a Markov chain, and since the process progresses in discrete time, it also satisfies the strong Markov property and hence

\[
(I_{k+1}^k, \tilde{X}_k) = (Y_{\tau(k)+1}, Y_{\tau(k)}, U_{\tau(k)}, X_{\tau(k)-1}, U_{\tau(k)-1})
\]

is a Markov chain. Now we show that it has a unique invariant measure under the assumption that the state process \(X_t\) admits a unique invariant measure under the exploring policy \(\gamma\).

We will denote the stationary distribution of \(X_t\) by \(\hat{\pi}\). Recalling that at the stopping time \(\tau(k)\) we have, \(Y_{\tau(k)} = y_1, Y_{\tau(k)-1} = y_0, U_{\tau(k)-1} = u_0, U_{\tau(k)} = u_1\), we can write:

\[
\lim_{k \to \infty} Pr(Y_{\tau(k)+1} = y_2, Y_{\tau(k)} = y_1, U_{\tau(k)} = u_1, X_{\tau(k)-1} = x_0, U_{\tau(k)-1} = u_0)
\]

\[
= \mathbb{1}_{\{Y_{\tau(k)} = y_1, U_{\tau(k)-1} = u_0, U_{\tau(k)} = u_1\}} P(y_2 | y_1, u_1, u_0, x_0) P_{\hat{\pi}}(x_0 | y_0)
\]
where

\[
P^\pi(x_0|y_0) := \lim_{k \to \infty} Pr(X_{\tau(k)-1} = x_0|Y_{\tau(k)-1} = y_0) = \frac{O(y_0|x_0)\hat{\pi}(dx_0)}{\int_{x_0} O(y_0|x'_0)\hat{\pi}(dx'_0)}.
\]

Then, for any measurable function \(f\) and for \(I = (y_1, y_0, u_0)\) and \(u = u_1\)

\[
\lim_{k \to \infty} \frac{1}{k} \sum_{k' = 0}^{k-1} f(I_{k'}, \hat{X}_{k'}) = \int_X \sum_{Y} f(y_2, y_1, u_1, x_0, u_0) P(y_2|y_1, u_1, u_0, x_0) P^\pi(dx_0|y_0).
\]

In particular, we have that

\[
\lim_{k \to \infty} \frac{1}{k} \sum_{k' = 0}^{k-1} V^*(I_{k'}) = \int_X \sum_{Y} V^*(y_2, y_1, u_1) P(y_2|y_1, u_1, u_0, x_0) P^\pi(dx_0|y_0)
\]

\[
= \sum_{I_1} V^*(I_1) P^*(I_1|I, u)
\]

where \(P^\pi(y_2|y_1, y_0, u_1, u_0)\) is the distribution of \(y_2\) when the \(x_0\)'s marginal distribution is given by \(\hat{\pi}\) and

\[
P^*(I_1 = (y_2, y'_1, u'_1)|I, u) := 1_{\{y'_1 = y_1, u'_1 = u_1\}} P^\pi(y_2|y_1, y_0, u_1, u_0)
\]

as defined in (6.14) and (6.15).

Using similar arguments, one can also show that for \(I = (y_1, y_0, u_0)\) and \(u = u_1\)

\[
\lim_{k \to \infty} \frac{1}{k} \sum_{k' = 0}^{k-1} C_{k'}(I, u) = \int_X c(x_1, u_1) P^\pi(dx_1|y_1, y_0, u_0)
\]

\[
= C^*(y_1, y_0, u_0, u_1) = C^*(I, u).
\]
Thus, we have that

\[ v_{k+1}(I, u) = \frac{1}{k} \sum_{k'=0}^{k} r_{k'}^{*}(I, u) \rightarrow 0 \]

almost surely for all \((I, u)\).
Chapter 7

Conclusion and Future Problems

In this thesis, we have focused on two main problems: (i) continuity and robustness properties of discrete time stochastic control problems with respect to transition kernels of the model; (ii) approximations and learning methods for partially observed Markov decision processes.

We have studied the robustness problem under the infinite horizon discounted cost criterion for fully observed and partially observed controlled processes. We have investigated the same problem under the infinite horizon average cost criterion for fully observed systems. For convergence of transition kernels, we have used weak convergence, setwise convergence and convergence under total variation distance. We have found that the continuity and the robustness can be guaranteed under the total variation convergence with mild assumptions, but for the weak and setwise convergence, additional regularity properties are needed for the continuity and the robustness. We have applied our findings to empirically constructed transition models where weak convergence is usually the natural convergence notion to work with.

For approximation methods in partially observed Markov processes, we have focused
on approximate models constructed using finite window history variables. We have provided two different techniques for both of which we have established error bounds relating the approximation performance to the window size, and we have established geometric convergence rates under controlled filter stability conditions. We have provided a comparison for two techniques where the first technique has lower error rates but the second technique results in a greater practical advantage with coarser error bounds.

Finally, we have studied the Q learning problem for partially observed MDPs using finite window information variables, and we have showed that the approximate Q function iterations converge to a fixed point equation that solves the optimality equation for the approximate belief MDP provided in POMDP approximation chapter. Consequentially, we have established error bounds for performance of learned policies. Furthermore, we have studied a Q learning algorithm for fully observed MDPs with continuous state spaces by discretizing the state space, and we have viewed the problem as a POMDP with a quantizer channel, where we have proved the near optimality of the learned policies.

For infinite horizon average cost problems, we have observed that the ergodicity of the state process is crucial for the analysis. Hence, for partially observed models, one needs to establish stability and ergodicity results for the filter process in order to analyze the infinite horizon average cost criterion for POMDPs (some related results include [19, 21, 20, 22]). Lack of such results has limited us to study the approximation results for POMDPs and robustness problem for partially observed systems under the average cost criterion. Hence, a possible research direction is the study of ergodicity properties of controlled filter processes and application of the results to the problems investigated in this thesis for POMDPs under the infinite horizon average cost criterion.

For approximation and learning problems in POMDPs, we have considered models
with finite observation spaces. Another immediate research direction is then to study the same problems for continuous observation spaces via discretizing the observation space and using the finite subset of the observation space to construct further approximations. Even for finite observation spaces, the methods we have used in this thesis, consider every element in the finite observation set. For computational efficiency, one can also construct approximation methods that make use of the observation realizations of POMDPs that are likely to be observed more often.

For both possible research directions, the analysis developed in this thesis will be applicable.
Bibliography


