OPTIMAL ZERO-DELAY TRANSMISSION OF MARKOV SOURCES: REINFORCEMENT LEARNING AND APPROXIMATIONS

by

LIAM CREGG

A thesis submitted to the
Department of Mathematics and Statistics
in conformity with the requirements for
the degree of Master of Applied Science

Queen's University
Kingston, Ontario, Canada
July 2024

Abstract

We study the problem of zero-delay coding for the transmission a Markov source over a noisy channel with feedback and present rigorous finite model approximations and reinforcement learning solutions which are guaranteed to achieve near-optimality. To this end, we formulate the problem as a Markov decision process (MDP) where the state is a probability-measure valued predictor/belief and the actions are quantizer maps. This MDP formulation has been used to show the optimality of certain classes of encoder policies in prior work. Despite such an analytical approach in determining optimal policies, their computation is prohibitively complex due to the uncountable nature of the constructed state space and the lack of minorization or strong ergodicity results which are commonly assumed for average cost optimal stochastic control. These challenges invite rigorous reinforcement learning methods, which entail several open questions addressed in our paper. We present two complementary approaches for this problem. In the first approach, we approximate the set of all beliefs by a finite set and use nearest-neighbor quantization to obtain a finite state MDP, whose optimal policies become near-optimal for the original MDP as the quantization becomes arbitrarily fine. In the second approach, a sliding finite window of channel outputs and quantizers together with a prior belief state serve as the state of the MDP. We then approximate this state by marginalizing over all possible beliefs, so that our policies only use the sliding finite window term to encode the source. Under an appropriate notion of predictor stability, we show that such policies are near-optimal for the zero-delay coding problem as the window length increases. We give sufficient conditions for predictor stability to hold. For each scheme, we propose a reinforcement learning algorithm to compute near-optimal policies. We provide a detailed comparison of the two coding policies in terms of their approximation bounds and reinforcement learning implementation, in terms of their performance, as well as conditions for reinforcement learning convergence to near-optimality. We include key differences between the noisy and noiseless channel cases, as well as supporting simulation results.

Co-Authorship

The following are colleagues who collaborated on results within this thesis (excluding supervisors Serdar Yüksel and Fady Alajaji):

Tamás Linder: The content of Chapter 2 is largely based on results from [1], where Tamás Linder is a co-author.

Acknowledgements

Firstly, I want to thank my supervisors Serdar Yüksel and Fady Alajaji, for their continued guidance, kindness, and patience. I have learned so much from them, both on an academic and personal level.

Thank you to my parents, for their unending love, and for believing in me. They have always fostered an environment where I could pursue my dreams, and I would not be where I am without their support.

I would also like to thank Tamás Linder for his collaboration on work within this thesis, as well as Yanglei Song and Brad Rodgers for taking the time to be on my thesis committee.

Finally, thank you Emmett, for always reminding me that two heads are better than one.

Contents

Chapter 1

Introduction

1.1 Motivation

In this thesis, we study the zero-delay coding problem, in which we wish to encode and transmit an information source at a fixed rate over a noisy channel with feedback and without delay, while minimizing the expected distortion at the receiver. In particular, we study the case where the information source is Markov and the noisy channel is memoryless. The zero-delay restriction is of practical relevance in many applications, including live-streaming and real-time sensor networks, or more generally any problem in which one wishes to communicate information quickly over a noisy channel. However, this restriction means that classical Shannon-theoretic methods [2], which require collecting large sequences of source symbols and compressing them at once, are not viable as they induce a large delay. While there exist several zero-delay coding algorithms which perform well in practice, they are usually heuristic and lack rigorous proofs of optimality. Our emphasis in this thesis is the development of zero-delay codes which are guaranteed to perform optimally or near-optimally.

Although zero-delay coding is an information-theoretic problem, several results in the literature have had success in using tools from the theory of stochastic control; this has yielded important theoretical results on the structure and existence of optimal codes. However, this approach has so far lacked effective algorithms for computing these optimal codes, perhaps due to the absence of the necessary results from the stochastic control literature.

Recently, there have been several results which generalize algorithms from stochastic control to settings which include the zero-delay coding problem. It is then natural to revisit this problem from a stochastic control perspective and attempt to apply these new results to obtain concrete algorithms for the zero-delay coding problem.

1.2 Problem Setup

Notation: In general, we will denote random variables by capital letters and their realizations by lowercase letters. There are a few exceptions to this; in particular we will always use lowercase π and uppercase Q in order to avoid a conflict of notation with existing results in the literature. It will be clear from the context for these variables whether we are referring to a random variable or its realization. To denote the set of probability measures over a measurable space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, we use $\mathcal{P}(\mathcal{X})$, and to denote a contiguous tuple of random variables $(X_0, X_1, \ldots, X_n)$ we will use the notation $X_{[0,n]}$ (and its realization by $x_{[0,n]}$). Probabilities and expectations will be denoted by P and $\mathbf{E}$, respectively. When the relevant distributions depend on some parameters, we include these in the superscript and/or subscript. For probabilities involving finite spaces, we will often use the shorthand $P(y_t|x_t) = P(Y_t = y_t|X_t = x_t)$, or simply $P(y|x)$, when the time index is not important. Also note that, even for a finite set $\mathcal{Y}$, we sometimes write for consistency of notation $\sum_{\mathcal{Y}} f(y)P(y|x) = \int_{\mathcal{Y}} f(y)P(dy|x)$, where we use the counting measure over $\mathcal{Y}$.

Describing our problem setup, let our information source be a time-homogeneous Markov process $(X_t)_{t\geq 0}$ taking values in $\mathcal{X}$, which we assume is finite. At each time $t \geq 0$, we wish to encode X_t as some channel input M_t and send it over a noisy channel, which yields output M_t'. The channel input and output alphabets will be denoted by $\mathcal{M}$ and $\mathcal{M}'$, respectively. We then wish to decode M_t' into some reproduction symbol $\hat{X}_t$, which takes values in the finite set $\hat{\mathcal{X}}$. The block diagram of the system is given in Figure 1.1 (note that we allow the encoder to have access to (noiseless) feedback from the channel).

We assume that the source $(X_t)_{t\geq 0}$ is irreducible and aperiodic. Accordingly, it admits a unique invariant measure, which we denote by ζ. We will denote the transition matrix of $(X_t)_{t\geq 0}$ by $T(x'|x) \coloneqq P(X_{t+1} = x'|X_t = x)$.

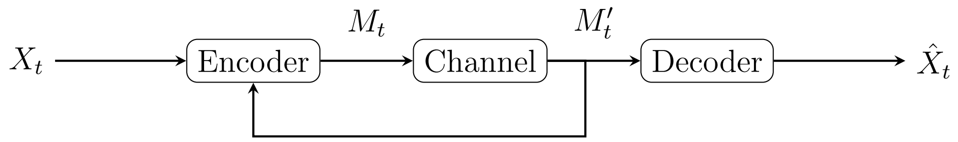

Figure 1.1: Source-channel coding with feedback

Let $X_0 \sim \pi_0$ (we will also call π_0 the *prior*). We also assume our channel is *memoryless* in the sense that M'_t is conditionally independent of $(X_{[0,t-1]}, M_{[0,t-1]}, M'_{[0,t-1]})$ given M_t and the channel is time invariant, and thus can be fully described by the transition matrix $O(m'|m) \coloneqq P(M'_t = m'|M_t = m)$.

Our admissible encoder and decoder have the following form: consider sequences of functions $\gamma^e \coloneqq (\gamma^e_t)_{t\geq 0}$, which we call the encoder policy, and $\gamma^d \coloneqq (\gamma^d_t)_{t\geq 0}$, which we call the decoder policy. In addition to the current source symbol, the encoder has access to all past source symbols and channel inputs, and all past channel outputs in the form of feedback. In addition to the current channel output, the decoder has access to all previous channel outputs. That is, $(\gamma^e_t)_{t\geq 0}$ and $(\gamma^d_t)_{t\geq 0}$ are such that

$$\begin{aligned} \gamma^e_t : \mathcal{X}^{t+1} \times \mathcal{M}^t \times (\mathcal{M}')^t &\to \mathcal{M} & \qquad \gamma^d_t : (\mathcal{M}')^{t+1} &\to \hat{\mathcal{X}} \\ (X_{[0,t]}, M_{[0,t-1]}, M'_{[0,t-1]}) &\mapsto M_t & M'_{[0,t]} &\mapsto \hat{X}_t. \end{aligned}$$

We consider two performance criteria for the zero-delay coding problem. We wish to find encoder and decoder policies such that one of the following distortion quantities is minimized: the discounted distortion,

$$\mathbf{E}_{\pi_0}^{\gamma^e,\gamma^d}\left[\sum_{t=0}^{\infty}\beta^t d(X_t,\hat{X}_t)\right], \tag{1.1}$$

or the average distortion,

$$\limsup_{T\to\infty}\mathbf{E}_{\pi_0}^{\gamma^e,\gamma^d}\left[\frac{1}{T}\sum_{t=0}^{T-1}d(X_t,\hat{X}_t)\right], \tag{1.2}$$

where $d:\mathcal{X}\times\hat{\mathcal{X}}\to\mathbb{R}_+$ is a given distortion function and $\beta\in(0,1)$ is a given discount factor. We use $\mathbf{E}_{\pi_0}^{\gamma^e,\gamma^d}$ and $P_{\pi_0}^{\gamma^e,\gamma^d}$ to denote expectations (respectively, probabilities) under encoder policy γ^e, decoder policy γ^d, and prior π_0, noting that these parameters induce a distribution on $\mathcal{X}^{\mathbb{Z}_+}\times\hat{\mathcal{X}}^{\mathbb{Z}_+}$.

We refer to the minimization of (1.1) as the discounted distortion problem and of (1.2) as the average distortion problem. Note that for a fixed encoder policy γ^e, it is straightforward to show that the optimal decoder policy, for all $t\geq 0$, is given by

$$\gamma_t^{d*}(M'_{[0,t]})=\operatorname*{argmin}_{\hat{x}\in\hat{\mathcal{X}}}\mathbf{E}_{\pi_0}^{\gamma^e}\left[d(X_t,\hat{x})|M'_{[0,t]}\right]. \tag{1.3}$$

Accordingly, we assume that we use an optimal decoder policy for a given encoder policy. With an abuse of notation, we then denote $\gamma:=\gamma^e$ and denote by Γ the set of all encoder policies. We can then write (1.1) and (1.2) as

$$J_\beta(\pi_0,\gamma) := \mathbf{E}_{\pi_0}^{\gamma}\left[\sum_{t=0}^{\infty}\beta^t d(X_t,\hat{X}_t)\right] \tag{1.4}$$

and

$$J(\pi_0,\gamma) := \limsup_{T\to\infty}\mathbf{E}_{\pi_0}^{\gamma}\left[\frac{1}{T}\sum_{t=0}^{T-1} d(X_t,\hat{X}_t)\right], \tag{1.5}$$

and the optimal respective costs by

$$J_\beta^*(\pi_0) := \inf_{\gamma\in\Gamma} J_\beta(\pi_0,\gamma) \tag{1.6}$$

and

$$J^*(\pi_0) := \inf_{\gamma\in\Gamma} J(\pi_0,\gamma). \tag{1.7}$$

We will also consider policies which obtain the above infima within some threshold $\epsilon > 0$; we say that a set of policies $\{\gamma\}$ depending on some parameter set is *near-optimal* for the discounted distortion problem (respectively, average distortion problem) if for any $\epsilon > 0$, there is some choice of parameters such that the resulting policy γ satisfies $J_\beta(\pi_0,\gamma) \leq J_\beta^*(\pi_0) + \epsilon$ (respectively, $J(\pi_0,\gamma) \leq J^*(\pi_0) + \epsilon$).

Note that in the zero-delay coding problem, we are usually concerned with the average distortion problem. However, we will show that as $\beta \to 1$, a policy that is near-optimal for the discounted distortion problem is also near-optimal for the average distortion problem.

1.3 Literature Review

From an information-theoretic perspective, several strategies have been used to approach this problem, including mutual information constraints, entropy coding, and Shannon lower bounding techniques. Studies to this end include [3]–[5]. Within the context of linear systems, [6]–[10] use sequential rate-distortion theory. Some of these works give applicable codes for zero-delay coding for Gaussian sources over additive-noise Gaussian channels, and some give upper and/or lower performance bounds; see also [11] and [12] for further studies.

Furthermore, learning theoretic methods have attracted significant interest in source-channel coding theory both in the classical literature and the recent literature, see for example [13]–[15] for the noiseless channel (quantization) case among several classical results, although usually restricted to independent and identically distributed (i.i.d.) sources. We note that our results are directly applicable for i.i.d. sources as well, since an optimal zero-delay code for an i.i.d. source is a memoryless code [16]–[18] (see also, for related discussions in a different causal coding context [19]–[22]).

More recently, deep learning is employed to construct powerful joint source-channel codes (see [23]–[26]), and reinforcement learning is used as a tool to estimate feedback capacity in [27], [28]. Although effective in practice, these machine learning methods are generally experimental and do not provide a formal proof of convergence or optimality. Conversely, our reinforcement learning approach will be rigorously shown to converge to near-optimality.

There have been several studies about the zero-delay coding problem using stochastic control techniques. In particular, [16], [17], [29] consider Markov sources with finite alphabets and finite time horizons and show optimality of structured classes of policies. Similar optimality and existence results are presented for infinite time horizons in [18] (with feedback) and [30] (without feedback). The continuous-alphabet infinite-horizon case is examined in [31], although only over a noiseless channel. These results often rely on formulating the problem as a Markov decision process (MDP) in order to utilize existing results from stochastic control theory, such as dynamic programming and value iteration methods (see [32], [33] for detailed information on such methods). However, in the formulation of the MDP, these results utilize a state space that is probability measure-valued (this state is often called the "predictor" in the literature) and an action space involving quantizers. These spaces are computationally difficult to work with, both in terms of complexity and implementation. Thus, while numerous existence and structural results have been established for this problem, the explicit development of effective coding schemes for a given zero-delay coding problem is still an open problem.

We will see in our development of the sliding finite window scheme that our method bears some resemblance to a trellis encoding scheme. Trellis codes (see e.g., [34]–[37]) use a sliding finite window combined with a tree-search algorithm to determine the optimal channel input, based on dynamic programming principles, while using an optimal filter at the decoder (note that trellis source coding generally performs the tree search at the encoder, while trellis channel coding generally performs it at the decoder [35]). Several joint source-channel coding theorems have been proven for trellis codes, including [36], which shows that trellis codes become optimal as the window length becomes large. Although trellis codes are not in general suitable for a zero-delay coding scheme (for the same reason as block codes), our encoder/decoder scheme may be seen in some sense as a modification of trellis codes to the zero-delay setting.

1.4 Markov Decision Processes

As our analysis will rely on the existing literature for MDPs, we give an introduction to the topic here and provide some important supporting results.

Definition 1.4.1. We define a *Markov decision process* (MDP) as a 4-tuple $(\mathcal{Z}, \mathcal{U}, P, c)$, where:

1. $\mathcal{Z}$ is the *state space*, which we assume is Polish (a Borel subset of a complete, separable metric space).

2. $\mathcal{U}$ is the *action space*, also Polish.

3. $P : \mathcal{Z} \times \mathcal{U} \to \mathcal{P}(\mathcal{Z})$ is the *transition kernel*, such that $(z, u) \mapsto P(dz'|z, u)$.

4. $c : \mathcal{Z} \times \mathcal{U} \to [0, \infty)$ is the *cost function.*

Given $Z_0 = z_0 \in \mathcal{Z}$, the objective is to minimize

$$J_\beta(z_0, f) := \mathbf{E}_{z_0}^f \left[\sum_{t=0}^{\infty} \beta^t c(Z_t, U_t) \right],$$

which we call the discounted cost problem, or

$$J(z_0, f) := \limsup_{T\to\infty} \mathbf{E}_{z_0}^f \left[\frac{1}{T} \sum_{t=0}^{T-1} c(Z_t, U_t) \right],$$

which we call the average cost problem, over all f, where $f = (f_t)_{t\geq 0}$ and $U_t = f_t(Z_{[0,t]}, U_{[0,t-1]})$.

The following is a classical result from the stochastic control literature

Theorem 1.4.2. *[32, Theorem 4.2.3] Let $\mathcal{Z}$ and $\mathcal{U}$ be finite, and define the* discounted cost optimality equation *(DCOE) as*

$$J_\beta^*(z) = \min_u \left\{ c(z, u) + \beta \sum_{z_1} J_\beta^*(z_1) P(z_1|z, u) \right\}.$$

A function satisfies the DCOE if and only if it is the optimal discounted cost; i.e. $J_\beta^*(z) = \inf_f J_\beta(z, f)$.

A Note on MDP Notation

The notation in the following sections can get intricate, so we first introduce some additional notation to be used in the context of MDPs. Some of the concise discussion below will be expanded upon and made more specific in the following sections.

1. When discussing *approximations* of an MDP state we use a caret symbol. For example, we use $\hat{Z}_t$ to denote an approximation of Z_t. Accordingly we use $\hat{f}$ to denote a policy that maps $\hat{Z}_t$ to U_t and we use $\hat{J}_\beta(\hat{Z}_0, \hat{f})$ to be some appropriately defined discounted cost under that policy. Furthermore, when the approximation depends on some parameter N, we denote such a policy by $\hat{f}_N$; when used in this way, we assume the policy is stationary, and thus the subscript N should not be confused with a time index.

2. To denote *optimality* we use an asterisk symbol. For example, we denote $J_\beta^*(z_0) := \inf_f J_\beta(z_0, f)$, and we denote a policy achieving this infimum by f^*.

3. Finally, when we *extend* an approximation back to the original state space, we use a tilde symbol. For example, if we wished to take $\hat{f}$ and appropriately modify it to take Z_t as an input rather than $\hat{Z}_t$, we would call the resulting policy $\widetilde{f}$.

Note that this notation is presented in order of operation. For example, $\hat{J}_\beta^*(\hat{z}_0)$ should be taken as the following: first we take an approximation of the MDP, then find the optimal discounted cost of this approximation. It is not the approximation of the optimal discounted cost $J_\beta^*(z_0)$, although under some conditions to be presented later it may be interpreted as such.

Accordingly, $\widetilde{J}^*_\beta(z_0)$ should be taken as the extension of $\hat{J}^*_\beta(\hat{z}_0)$ (as a function of $\hat{z}_0$) to all of $\mathcal{Z}$, not as the optimal discounted cost for some "extended" MDP.

1.4.1 MDP Approximation

We first recall some results from [38] on approximation of MDPs. The following is an important property of transition kernels.

Definition 1.4.3. We say a transition kernel $P(dz'|z,u)$ is weakly continuous if for all continuous and bounded $f : \mathcal{Z} \to \mathbb{R}$

$$\int f(z')P(dz'|z,u)$$

is continuous in (z,u).

We note that MDPs with weakly continuous transition kernels as above are often called *weak Feller*. We assume the following:

Assumption 1.4.4. *(i) The cost function c is continuous and bounded.*

(ii) The transition kernel P is weakly continuous.

(iii) $\mathcal{Z}$ and $\mathcal{U}$ are compact.

Let $d_\mathcal{Z}$ be the metric on $\mathcal{Z}$. By compactness, there exists a sequence of finite grids $\mathcal{Z}_N = \{\hat{z}_{N,1}, \ldots, \hat{z}_{N,m_N}\} \subset \mathcal{Z}$ such that

$$\max_{z \in \mathcal{Z}} \min_{i=1,\ldots,m_N} d_\mathcal{Z}(z, \hat{z}_{N,i}) \to 0$$

as $N \to \infty$. Here N may be interpreted as a resolution parameter of the approximation, which we allow to become arbitrarily high. Then, recalling the notation in Section 1.4, we define $\hat{z}$ as

$$\hat{z} \coloneqq \operatorname*{argmin}_{z' \in \mathcal{Z}_N} d_{\mathcal{Z}}(z, z').$$

That is, $\hat{z}$ is the nearest neighbor of z in $\mathcal{Z}_N$, where ties are broken so that the map $z \mapsto \hat{z}$ is measurable. This map induces a partition $\{B_{N,i}\}_{i=1}^{m_N}$ of $\mathcal{Z}$ where $B_{N,i} = \{z \in \mathcal{Z} : q(z) = \hat{z}_{N,i}\}$, where we have used q to denote the nearest neighbor map.

Finally, let $(\nu_N)_{N \geq 0} \subset \mathcal{P}(\mathcal{Z})$ be such that $\nu_N(B_{N,i}) > 0$ for all N, i and let

$$\nu_{N,i}(\cdot) = \frac{\nu_N(\cdot)}{\nu_N(B_{N,i})}.$$

Then we approximate the MDP $(\mathcal{Z}, \mathcal{U}, P, c)$ with a new MDP $(\mathcal{Z}_N, \mathcal{U}, P_N, c_N)$, where P_N and c_N are the averages of P and c over the appropriate $B_{N,i}$ with respect to $\nu_{N,i}$. That is,

$$\begin{aligned} P_N(\hat{z}_{N,j} | \hat{z}_{N,i}, u) &= \int_{B_{N,i}} P(B_{N,j} | z, u) \nu_{N,i}(dz) \\ c_N(\hat{z}_{N,i}, u) &= \int_{B_{N,i}} c(z, u) \nu_{N,i}(dz). \end{aligned} \tag{1.8}$$

Note that the new MDP has a finite state space. We denote this new MDP by $\text{MDP}_N \coloneqq (\mathcal{Z}_N, \mathcal{U}, P_N, c_N)$. Finally, note that we can extend any policy $\hat{f}_N$ defined for MDP_N to the original MDP by making it constant over the $B_{N,i}$. Again recalling our notation in Section 1.4, we denote this extension by $\widetilde{f}_N$; that is,

$$\widetilde{f}_N(z) = f_N(\hat{z}).$$

The following is a key result by [38], which states that policies which are optimal for MDP$_N$, when appropriately extended, become near-optimal for the true MDP as N gets large. According to the notation in Section 1.4, we denote the optimal policy for MDP$_N$ by $\hat{f}_N^*$ and its extension to all of $\mathcal{Z}$ by $\widetilde{f}_N^*$

Theorem 1.4.5. *[38, Theorem 4.3] Let Assumption 1.4.4 hold. Then for all $z_0 \in \mathcal{Z}$ and $\beta \in (0,1)$,*

$$\lim_{N\to\infty} |J_\beta(z_0, \widetilde{f}_N^*) - J_\beta^*(z_0)| = 0.$$

1.4.2 Q-learning Convergence

We now recall some recent results from [39] regarding the convergence of certain "Q-learning" iterations. Let $(S_t)_{t\geq 0}$, $(U_t)_{t\geq 0}$, and $(C_t)_{t\geq 0}$ be $\mathcal{S}$-valued, $\mathcal{U}$-valued, and $\mathbb{R}$-valued stochastic processes, respectively. Define $V_t : \mathcal{S} \times \mathcal{U} \to \mathbb{R}$ by

$$V_{t+1}(S_t, U_t) = (1 - \alpha_t(S_t, U_t))V_t(S_t, U_t) + \alpha_t(S_t, U_t)\left(C_t + \beta \min_{u\in\mathcal{U}} V_t(S_{t+1}, u)\right)$$
$$V_{t+1}(s, u) = V_t(s, u) \quad \text{for all } (s, u) \neq (S_t, U_t), \tag{1.9}$$

where

$$\alpha_t(s,u) = \frac{1}{1 + \sum_{k=0}^{t} \mathbf{1}(S_k = s, U_k = u)}.$$

The following ergodicity assumptions are sufficient for the convergence of V_t.

Assumption 1.4.6. *[39, Assumption 2.2] the process $(S_{t+1}, S_t, U_t, C_t)_{t\geq 0}$ is such that almost surely,*

(i) For all (s,u), $\sum_{t\geq 0} \alpha_t(s,u) = \infty$.

(ii)

$$\frac{\sum_{k=0}^{t} C_k \mathbf{1}(S_k = s, U_k = u)}{\sum_{k=0}^{t} \mathbf{1}(S_k = s, U_k = u)} \to C^*(s,u)$$

for some $C^* : \mathcal{S} \times \mathcal{U} \to \mathbb{R}$.

(iii)

$$\frac{\sum_{k=0}^{t} f(S_{k+1}) \mathbf{1}(S_k = s, U_k = u)}{\sum_{k=0}^{t} \mathbf{1}(S_k = s, U_k = u)} \to \int f(s_1) P^*(ds_1|s,u)$$

for any $f : \mathcal{S} \to \mathbb{R}$ and some $P^* : \mathcal{S} \times \mathcal{U} \to \mathcal{P}(\mathcal{S})$.

Theorem 1.4.7. *[39, Theorem 2.1] Under Assumption 1.4.6, for every (s,u), $V_t(s,u)$ converges to $V^*(s,u)$ satisfying*

$$V^*(s,u) = C^*(s,u) + \beta \sum_{\mathcal{S}} \min_u V^*(s_1,u) P^*(s_1|s,u) \tag{1.10}$$

Remark 1.4.8. By taking the minimum of V^* over $\mathcal{U}$, we recapture exactly the DCOE from 1.4.2, and thus the policy $\gamma^*(s) := \operatorname{argmin}_u V^*(s,u)$ is optimal (in discounted cost) for the MDP defined by $(\mathcal{S}, \mathcal{U}, P^*, C^*)$. Note that we do *not* require that S_t is actually distributed according to $P^*(\cdot|s,u)$, only that its long-term sample average converges to P^* in the sense of (iii).

1.4.3 Zero-delay Coding as an MDP

Returning to our zero-delay coding problem, for fixed $x_{[0,t-1]}$, $m_{[0,t-1]}$ and $m'_{[0,t-1]}$, consider the function

$$\gamma(\cdot, x_{[0,t-1]}, m_{[0,t-1]}, m'_{[0,t-1]}) : \mathcal{X} \to \mathcal{M}.$$

Such a function (that is, a mapping from $\mathcal{X}$ to $\mathcal{M}$) is called a *quantizer*. We denote the set of all quantizers by $\mathcal{Q}$. Thus we can view a policy γ as selecting a quantizer $Q_t \in \mathcal{Q}$ based on the information $(X_{[0,t-1]}, M_{[0,t-1]}, M'_{[0,t-1]})$, then generating the channel input M_t as $Q_t(X_t)$, as in [40].

Recall that we used $O(m'|m)$ to denote our channel transition kernel. Let $O_Q(m'|x)$ denote the kernel induced by a quantizer $Q \in \mathcal{Q}$; that is, $O_Q(m'|x) = O(m'|Q(x))$. Now let $\psi \in \mathcal{P}(\mathcal{M}')$ be such that $O_Q(\cdot|x) \ll \psi$ for all $x \in \mathcal{X}, Q \in \mathcal{Q}$, where we use "$\ll$" to denote absolute continuity (that is, $\psi(B) = 0 \implies O_Q(B|x) = 0$ for any Borel $B \subset \mathcal{M}'$). Since $\mathcal{M}'$ is finite in our setup, we will take ψ to be the uniform measure on $\mathcal{M}'$, but note that such measures also exists in uncountable setups for most practical channels. Then let $g_Q(x, m') := \frac{dO_Q}{d\psi}(x, m')$ be the Radon-Nikodym derivative of O_Q with respect to ψ. In particular for a uniform ψ, we have $g_Q(x, m') = |\mathcal{M}'|\, O_Q(m'|x)$

Also, let $\pi_t, \overline{\pi}_t \in \mathcal{P}(\mathcal{X})$ be defined as

$$\pi_t(\cdot) = P^{\gamma}_{\pi_0}(X_t \in \cdot | M'_{[0,t-1]}) \tag{1.11}$$

$$\overline{\pi}_t(\cdot) = P^{\gamma}_{\pi_0}(X_t \in \cdot | M'_{[0,t]}), \tag{1.12}$$

recalling that $X_0 \sim \pi_0$. We have dropped the policy γ for notational simplicity, but it should be noted that π_t and $\overline{\pi_t}$ are policy-dependent. In the literature, π_t is called the *predictor* and $\overline{\pi}_t$ is called the *filter.* Note that throughout the thesis, we will refer to the predictor as the belief, although in the POMDP literature the term belief is typically saved for filters. With a slight abuse of notation, we also let the source transition kernel T act as an operator on probability measures as follows:

$$\begin{aligned} T : \mathcal{P}(\mathcal{X}) &\to \mathcal{P}(\mathcal{X}) \\ \pi(x) &\mapsto \sum_{\mathcal{X}} T(x'|x)\pi(x). \end{aligned}$$

Then given π_0, the above measures can be computed in a recursive manner as follows (see [41, Proposition 3.2.5]).

$$\overline{\pi}_t(x) = \frac{g_{Q_t}(x, M'_t)\pi_t(x)}{\sum_{\mathcal{X}} g_{Q_t}(x, M'_t)\pi_t(x)},$$

$$\pi_{t+1} = T(\overline{\pi}_t). \tag{1.13}$$

We denote $N(m', Q) := \sum_{\mathcal{X}} g_Q(x, m')\pi_t(x)$. Note that $N(M'_t, Q_t)$ is non-zero $P^{\gamma}_{\pi_0}$ almost surely (a.s.). Thus inside of $P^{\gamma}_{\pi_0}$ expectations we assume $N(M'_t, Q_t)$ is non-zero.

Using the above update equations, one can compute π_t given $(M'_{[0,t-1]}, Q_{[0,t-1]})$, so that policies of the form $Q_t = \gamma_t(\pi_t)$ are valid. These policies form a special class.

Definition 1.4.9. [18] We say a policy $\gamma = \{\gamma_t\}_{t\geq 0}$ is of the *Walrand-Varaiya type* if, at time t, γ selects a quantizer $Q_t = \gamma_t(\pi_t)$ and M_t is generated as $M_t = Q_t(X_t)$. Such a policy is called *stationary* if it does not depend on t (that is, $\gamma_t = \overline{\gamma}$ for some $\overline{\gamma}$ and all $t \geq 0$). The set of all stationary Walrand-Varaiya policies is denoted by Γ_{WS}.

The following are key results, originally from Walrand and Varaiya [16] for a finite time horizon and extended to the infinite-horizon case in [18].

Proposition 1.4.10. *[18, Proposition 2] For any $\beta \in (0,1)$, there exists $\gamma^* \in \Gamma_{WS}$ that solves the discounted distortion problem (that is, it minimizes* (1.1)*) for all priors $\pi_0 \in \mathcal{P}(\mathcal{X})$.*

Proposition 1.4.11. *[18, Theorem 3] There exists $\gamma^* \in \Gamma_{WS}$ that solves the average distortion problem (that is, it minimizes* (1.2)*) for all priors $\pi_0 \in \mathcal{P}(\mathcal{X})$.*

Proposition 1.4.12. *Under any $\gamma \in \Gamma_{WS}$, the zero-delay coding problem is an MDP, where:*

1. $\mathcal{Z} = \mathcal{P}(\mathcal{X})$.
2. $\mathcal{U} = \mathcal{Q}$.
3. $P = P(\cdot|\pi, Q)$ *is induced by the update equations in* (1.13).
4. $c(\pi, Q) = \sum_{\mathcal{M}'} \min_{\hat{x}\in\hat{\mathcal{X}}} \sum_{\mathcal{X}} d(x,\hat{x}) O_Q(m'|x)\pi(x)$.

This follows directly from the update equations in (1.13) and the fact that, under any $\gamma \in \Gamma_{WS}$, π_t completely determines Q_t. The choice of c is due to the following result.

Lemma 1.4.13. *If an optimal decoder is used, the expected distortion at the encoder (that is, before sending M_t) is given by*

$$c(\pi_t, Q_t) = \sum_{\mathcal{M}'} \min_{\hat{x}\in\hat{\mathcal{X}}} \sum_{\mathcal{X}} d(x,\hat{x}) O_{Q_t}(m'|x)\pi_t(x). \tag{1.14}$$

Proof. Recall that, for a fixed Q_t, the optimal decoder γ^{d*} chooses $\hat{X}_t$ according to

$$\gamma_t^{d*}(M'_{[0,t]}) = \operatorname*{argmin}_{\hat{x}} \mathbf{E}_{\pi_0}^{\gamma}\left[d(X_t,\hat{x})|M'_{[0,t]}\right] = \operatorname*{argmin}_{\hat{x}} \sum_{x} d(x,\hat{x})\overline{\pi}_t(x).$$

By the update equations in (1.13), we have

$$\overline{\pi}_t(x) = \frac{g_{Q_t}(x, M'_t)\pi_t(x)}{N(M'_t, Q_t)},$$

so that at the decoder the expected distortion is given by

$$\min_{\hat{x}} \sum_x d(x,\hat{x}) \frac{g_{Q_t}(x, M'_t)\pi_t(x)}{N(M'_t, Q_t)}.$$

However, at the encoder we must take the further expectation over M'_t (conditioned on $M'_{[0,t-1]}$), since we do not yet have access to M'_t. Thus, at the encoder the expected distortion is

$$\begin{aligned}
&\sum_{m'} \min_{\hat{x}} \sum_x d(x,\hat{x}) \frac{g_{Q_t}(x, m')\pi_t(x)}{N(m', Q)} P^{\gamma}_{\pi_0}(m'|M'_{[0,t-1]}) \\
=&\sum_{m'} \min_{\hat{x}} \sum_x d(x,\hat{x}) \frac{g_{Q_t}(x, m')\pi_t(x)}{N(m', Q_t)} \sum_x \pi_t(x) O_{Q_t}(m'|x) \\
=&\sum_{m'} \min_{\hat{x}} \sum_x d(x,\hat{x}) O_{Q_t}(m'|x)\pi_t(x).
\end{aligned}$$

The first equality holds by marginalizing over X_t and using conditional probability rules, and the second by using the definitions of $g_Q(x, m')$ and $N(m', Q)$. □

By this lemma, we have that the expected distortion at the encoder (assuming an optimal decoder), satisfies for all $T \geq 1$,

$$\mathbf{E}^{\gamma}_{\pi_0}\left[\sum_{t=0}^{T-1} c(\pi_t, Q_t)\right] = \mathbf{E}^{\gamma}_{\pi_0}\left[\sum_{t=0}^{T-1} d(x_t, \hat{x}_t)\right].$$

Thus, this choice of c ensures that solving the MDP defined in Proposition 1.4.12 over all $\gamma \in \Gamma_{WS}$ (that is, minimizing $J_\beta(\pi_0, \gamma)$ or $J(\pi_0, \gamma)$) is equivalent to solving the zero-delay coding problem. Accordingly, we hereafter consider the discounted and average cost problems (rather than the discounted and average distortion problems). This allows us to use strategies from the literature of stochastic control; however, several complexities have been introduced:

- While the source alphabet $\mathcal{X}$ is finite, the state space of the MDP, $\mathcal{P}(\mathcal{X})$, is uncountable. Furthermore, while our source process $(X_t)_{t\geq 0}$ is irreducible and aperiodic (and hence has a unique invariant measure), there is no a priori reason for the MDP state process $(\pi_t)_{t\geq 0}$ to inherit these properties; in particular, irreducibility is too demanding.
- While we assume knowledge of the source transition kernel T, the calculation of the transition kernel $P(d\pi'|\pi, Q)$ is computationally demanding.

Thus even if one can approximate the MDP state space $\mathcal{P}(\mathcal{X})$ by some finite one, implementation of traditional MDP methods such as dynamic programming is difficult for this problem. This motivates the use of a reinforcement learning approach in which the calculation of these transition probabilities is unnecessary. Finally, although explicit computation of $P(d\pi'|\pi, Q)$ is difficult, the following key structural result was obtained in [40].

Lemma 1.4.14. *[40, Lemma 11] The transition kernel $P(d\pi'|\pi, Q)$ is weakly continuous (recall Definition 1.4.3).*

Here, we endow $\mathcal{P}(\mathcal{X})$ with the weak convergence topology and $\mathcal{Q}$ with the Young topology (see [40]). Alternatively, since $\mathcal{Q}$ is finite here the discrete topology would also suffice.

Finally, we conclude our MDP discussion by presenting a connection between the discounted cost problem and the average cost problem in the context of zero-delay coding, which was established in [1]. It implies that, for the zero-delay coding problem, a policy which is near-optimal for the discounted cost problem for β sufficiently close to 1, is also near-optimal for the average cost problem.

Theorem 1.4.15. *[1, Theorem 5] For every $\epsilon > 0$, there exists β' such that for all $\beta \in (\beta', 1)$, if $\gamma_\beta \in \Gamma_{WS}$ satisfies $J_\beta(\pi_0, \gamma_\beta) \leq J^*_\beta(\pi_0) + \delta$, then*

$$J(\pi_0, \gamma_\beta) \leq J^*(\pi_0) + \epsilon + (1 - \beta)\delta.$$

Remark 1.4.16. This result is important since from a zero-delay coding perspective, the objective is usually the average cost problem. However, the discounted cost problem is much easier to tackle from a reinforcement learning standpoint, and crucially will allow us to use Theorem 1.4.7. Accordingly, the majority of this thesis will address near-optimality for the discounted cost problem, and then near-optimality for the average cost follows from Theorem 1.4.15.

1.5 Filter and Predictor Stability

A key property that we use is *filter/predictor* stability (recall from Definition 1.11 that the predictor is given by π_t and the filter by $\overline{\pi}_t$).

Definition 1.5.1. The total variation distance between two probability measures μ, ν defined over $\mathcal{X}$ is given by

$$||\mu - \nu||_{TV} := \sup_{||f||_\infty \leq 1} \left| \int_{\mathcal{X}} f(x)\mu(dx) - \int_{\mathcal{X}} f(x)\nu(dx) \right|,$$

where the supremum is over all measurable real functions such that $||f||_\infty = \sup_{x \in \mathcal{X}} |f(x)| \leq 1$.

Note that the total variation distance is equivalent to the L_1 metric when $\mathcal{X}$ is finite. Recall the update equations in (1.13) and note that they are sensitive to the value of π_0. Accordingly, we use π_t^μ to denote the predictor when $\pi_0 = \mu$. The question of *predictor stability* is the following: under what conditions is $(\pi_t)_{t \geq 0}$ insensitive to its initialization, in the sense that $\lim_{t \to \infty} ||\pi_t^\mu - \pi_t^\nu||_{TV} = 0$ when π_t^μ and π_t^ν are updated with the same sequence $M'_{[0,t-1]}$? We will study several types of predictor stability in this thesis.

We can ask a similar question for the filter process $(\overline{\pi}_t)_{t \geq 0}$; in fact, the problem of filter stability (in various senses) is a classical problem in probability and statistics, where it is typically established in two ways: (i) The transition kernel of the underlying state is in some sense *sufficiently ergodic*, so that regardless of the observations, the filter process inherits this ergodicity and forgets its prior over time. (ii) The observations are in some sense *sufficiently informative*, so that, regardless of the prior, the filter process tracks the true state process. For a detailed review of these filter stability methods, see [42]. However, we will need slightly more general results in our case, since it is usually assumed in the filter stability problem that the observation kernel is time-invariant; here O_{Q_t} depends on Q_t and hence changes with time, and accordingly additional analysis is needed.

1.6 Contributions

In this thesis, we present two rigorous approximation methods to simplify the resulting MDP, and we use these approximations to obtain near-optimal coding schemes for the zero-delay coding problem via a reinforcement learning approach. We emphasize that we provide guaranteed approximation and convergence results. In particular, we build on methods used in [43] and [44], which were originally used to study partially observed Markov decision processes (POMDPs). The first is a discretization (or quantization) of the probability measure-valued state space using a nearest-neighbor map, and the second is based on an approximation of the probability measure using a sliding finite window of past observations.

The former method was used in [1] to study the noiseless channel (quantization) problem and similarly obtain near-optimal codes. In this work, we significantly extend those results to the noisy channel case, which requires some additional analysis, due to the fact that under the noiseless channel setup the filter/predictor process always admits a recurrent state under a uniform exploration policy, making the stochastic analysis on the ergodicity properties less demanding. More importantly, we introduce an alternative and more practical finite sliding window method, which was not covered in [1], and we present several mathematical and algorithmic results about its near-optimal performance.

For both methods, we analyze and rigorously establish the convergence of a simple reinforcement learning algorithm to obtain near-optimal codes for the original zero-delay coding problem, facilitated by the approximations. These two schemes are complementary: the quantization approach requires weaker assumptions, but comes at the cost of additional computational complexity, and sensitivity to initialization. In particular, the approximations are near-optimal as long as the source has a unique invariant distribution, but the resulting policy is only valid when the source starts from this invariant distribution. The sliding finite window approach has stricter conditions, but is very simple to implement. In particular, it requires Dobrushin coefficient conditions on the source and channel, but the resulting policy is less computationally complex and it is valid for any initial window (see Table 5.1).

1.7 Organization of Thesis

Chapter 2 is a preliminary chapter in which we cover the noiseless channel case; this chapter is based on the results in [1].[1] While not the main content of the thesis, this chapter will form the basis for the analysis presented in later chapters. We review the MDP formulation of the problem in this case, and present a quantization-based approximation scheme. We also give reinforcement learning results, which are greatly facilitated by the lack of channel noise.

In Chapter 3, we extend the results of Chapter 2 to the case of a noisy channel, which in particular requires additional analysis of the ergodic properties of the predictor. We present asymptotic near-optimality results under very weak assumptions.

In Chapter 4, we provide a new approximation scheme using a sliding finite window of past observations. We show a stronger exponential convergence to optimality, but at the cost of stricter assumptions on the source and channel. As supporting results, we extend several filter stability results in the literature to the zero-delay coding setup.

In Chapter 5 we present a detailed comparison of the approximation schemes in Chapters 3 and 4, both from a theoretical and implementation perspective. Finally, we provide conclusions and suggest future research directions in Chapter 6.

[1] This was performed in part when the author was a USRA student from May-August 2022.

Chapter 2

Preliminary Results: Belief-Quantization Based Coding for Noiseless Channels

2.1 Near-Optimality Results via a Finite MDP Approximation

In the noiseless channel case, our channel matrix $O(m'|m)$ is the identity matrix, and in this case our update equation in (1.13) simplifies to

$$\pi_{t+1}(x') = \frac{1}{\pi_t(Q_t^{-1}(M_t))} \sum_{x \in Q_t^{-1}(M_t)} T(x'|x)\pi_t(x), \tag{2.1}$$

and our cost given in (1.14) simplifies to

$$c(\pi, Q) = \sum_{m \in \mathcal{M}} \min_{\hat{x} \in \hat{\mathcal{X}}} \sum_{x \in Q^{-1}(m)} d(x, \hat{x}) \pi(x). \tag{2.2}$$

Note that feedback is not needed in this case, since $M_t = M_t'$ for all $t \geq 0$.

Following the MDP approximation scheme in Section 1.4.1, we approximate our state space $\mathcal{P}(\mathcal{X})$ by the following finite set. Given $N \in \mathbb{Z}_+$, define

$$\mathcal{P}_N(\mathcal{X}) = \left\{ \hat{\pi} \in \mathcal{P}(\mathcal{X}) : \hat{\pi} = \left[\frac{k_1}{N}, \ldots, \frac{k_{|\mathcal{X}|}}{N} \right], k_i = 0, \ldots, N, i = 1, \ldots, |\mathcal{X}| \right\}, \tag{2.3}$$

and given $\pi \in \mathcal{P}(\mathcal{X})$, let $\hat{\pi}$ be the nearest neighbor (in Euclidean distance) of π in $\mathcal{P}_N(\mathcal{X})$. We clearly have that $\max_\pi d(\pi, \hat{\pi}) \to 0$ as $N \to \infty$ (explicit calculations for this maximum are given in [45, Proposition 2] under several norms). Accordingly, define $P_N(\hat{\pi}_j | \hat{\pi}_i, Q)$ and $c_N(\hat{\pi}_i, Q)$ as

$$\begin{aligned} P_N(\hat{\pi}_j | \hat{\pi}_i, Q) &= \int_{B_i} P(B_j | \pi, Q) \nu_{N,i}(d\pi) \\ c_N(\hat{\pi}_i, Q) &= \int_{B_i} c(\pi, Q) \nu_{N,i}(d\pi), \end{aligned} \tag{2.4}$$

where B_j and B_i are the bins (under the nearest neighbor map) corresponding to $\hat{\pi}_j$ and $\hat{\pi}_i$, respectively. In order to apply Theorem 1.4.5 to $\text{MDP}_N := (\mathcal{P}_N(\mathcal{X}), \mathcal{Q}, P_N, c_N)$, we must show that Assumption 1.4.4 holds for the true zero-delay coding MDP given by $(\mathcal{P}(\mathcal{X}), \mathcal{Q}, P, c)$, which we formalize in the following lemma.

Lemma 2.1.1. *The zero-delay coding MDP* $(\mathcal{P}(\mathcal{X}), \mathcal{Q}, P, c)$ *meets Assumption 1.4.4.*

Proof. (i) holds by noting that $c(\pi, Q)$ in (2.2) is continuous in π and that $\mathcal{Q}$ is finite. Since $\mathcal{X}$ is finite, $\mathcal{P}(\mathcal{X})$ is compact and so we also have boundedness. (ii) follows from Lemma 1.4.14. Finally, (iii) holds since $\mathcal{X}$ and $\mathcal{M}$ are finite, which implies compactness of $\mathcal{P}(\mathcal{X})$ and finiteness of $\mathcal{Q}$. □

The following is then simply an application of Theorem 1.4.5 to the zero-delay coding MDP.

Corollary 2.1.2. *Let* $MDP_N = (\mathcal{P}_N(\mathcal{X}), \mathcal{Q}, P_N, c_N)$*. Let* $\hat{\gamma}_N^* \in \Gamma_{WS}$ *be optimal for* MDP_N*, and for any* π *let* $\widetilde{\gamma}_N^*(\pi) = \hat{\gamma}_N^*(\hat{\pi})$*, where* $\hat{\pi}$ *is the nearest neighbor of* π *in* $\mathcal{P}_N(\mathcal{X})$*. Then for all* $\pi_0 \in \mathcal{P}(\mathcal{X})$ *and* $\beta \in (0, 1)$*,*

$$\lim_{N\to\infty} |J_\beta(\pi_0, \widetilde{\gamma}_N^*) - J_\beta^*(\pi_0)| = 0.$$

That is, $\widetilde{\gamma}_N^*$ *is near-optimal for the zero-delay coding problem under the discounted distortion criterion.*

2.2 Q-learning and its Convergence to a Near-Optimal Finite MDP

We now present a reinforcement learning algorithm to compute the policy $\hat{\gamma}_N^*$ from the previous corollary, based on the results in Section 1.4.2. Consider the following algorithm for computing the sequences $(\hat{\pi}_t)_{t\geq 0}, (Q_t)_{t\geq 0}$, and $(C_t)_{t\geq 0}$.

Algorithm 1: Quantized Q-learning for noiseless channel

```
Require: initial distribution $\pi_0$, transition kernel $T$, quantizer set $\mathcal{Q}$
Sample $X_0 \sim \pi_0$
Choose $Q_0$ uniformly from $\mathcal{Q}$
$M_0 = Q_0(X_0)$
Compute $C_0 = c(\pi_0, Q_0)$ using (2.2)
for $t \geq 1$ do
    Compute $\pi_t$ using (2.1)
    Sample $X_t \sim T(\cdot|X_{t-1})$
    Choose $Q_t$ uniformly from $\mathcal{Q}$
    $M_t = Q_t(X_t)$
    Compute $C_t = c(\pi_t, Q_t)$ using (2.2)
```

Then consider the sequence $(V_t)_{t\geq 0}$, where $V_t : \mathcal{P}_N(\mathcal{X}) \times \mathcal{Q} \to \mathbb{R}_+$, defined by

$$V_{t+1}(\hat{\pi}_t, Q_t) = (1 - \alpha_t(\hat{\pi}_t, Q_t))V_t(\hat{\pi}_t, Q_t) + \alpha_t(\hat{\pi}_t, Q_t)\left(C_t + \beta \min_{Q \in \mathcal{Q}} V_t(\hat{\pi}_{t+1}, Q)\right)$$

$$V_{t+1}(\hat{\pi}, Q) = V_t(\hat{\pi}, Q) \quad \text{for all } (\hat{\pi}, Q) \neq (\hat{\pi}_t, Q_t), \tag{2.5}$$

where

$$\alpha_t(\hat{\pi}, Q) = \frac{1}{1 + \sum_{k=0}^{t} \mathbf{1}(\hat{\pi}_k = \hat{\pi}, Q_k = Q)}.$$

In order to apply Theorem 1.4.7, we need to show that Assumption 1.4.6 holds for the sequence $(S_{t+1}, S_t, U_t, C_t)_{t\geq 0} \coloneqq (\hat{\pi}_{t+1}, \hat{\pi}_t, Q_t, C_t)_{t\geq 0}$. The remainder of this section is dedicated to proving this result. Recall that $(\pi_t^{\mu})_{t\geq 0}$ is used to denote the predictor process with initialization $\pi_0 = \mu$.

Lemma 2.2.1. *Let Q_t be chosen uniformly and randomly for all $t \geq 0$, as in Algorithm 1. Then there exists an element $\pi^* \in \mathcal{P}(\mathcal{X})$ such that*

$$\tau \coloneqq \inf\{t \geq 0 : \pi_t = \pi^*\}$$

satisfies $P_{\pi_0}(\tau < \infty) = 1$ for all $\pi_0 \in \mathcal{P}(\mathcal{X})$.

Proof. Consider a quantizer $Q \in \mathcal{Q}$ such that for some $x \in \mathcal{X}$ and $m \in \mathcal{M}$, we have $Q^{-1}(m) = \{x\}$. That is, Q quantizes at least one element of $\mathcal{X}$ without any loss. By direct computation using (2.1), when $Q_t = Q$ and $X_t = x$, we have

$$\pi_{t+1}(x') = T(x'|x).$$

That is, π_{t+1} becomes the row of T corresponding to x, regardless of π_t. But since $\mathcal{X}$ and $\mathcal{Q}$ are finite, and since $(X_t)_{t\geq 0}$ is irreducible and aperiodic, the event $(X_t, Q_t) = (x, Q)$ for some $t \geq 0$ happens almost surely. Letting

$\pi^* = T(\cdot|x)$ in the lemma statement, the result follows. $\square$

Remark 2.2.2. Note that this lemma gives a very strong type of recurrence which crucially uses the fact that the channel is noiseless; in the general noisy case, such an element π^* is not guaranteed to exist.

Lemma 2.2.3. *Let Q_t be chosen uniformly and randomly for all $t \geq 0$, as in Algorithm 1. Then the following hold:*

(i) The predictor process $(\pi_t)_{t\geq 0}$ admits a unique invariant measure ϕ.

(ii) For any initialization $\pi_0 = \mu$ and for any measurable and bounded function $f : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$, we have

$$\frac{1}{T}\sum_{t=0}^{T-1} f(\pi_t^{\mu}) \to \int f d\phi$$

P^{μ} almost surely as $T \to \infty$.

Proof. To show (i), note that when Q_t is chosen randomly and independently of π_t, the induced transition kernel $P(d\pi'|\pi)$ becomes weakly continuous. Since every Markov process with a weakly continuous transition kernel on a compact state space admits an invariant measure [46, Theorem 7.2.3], $(\pi_t)_{t\geq 0}$ has an invariant measure. Thus, we are left with proving uniqueness.

Now suppose there exist two distinct invariant measures for $(\pi_t)_{t\geq 0}$. This implies (see for example [46, Lemma 2.2.3]) that there exist two mutually singular invariant measures ϕ_1, ϕ_2 and two disjoint sets $B_1, B_2 \subset \mathcal{P}(\mathcal{X})$ such that $\phi_1(B_1) = \phi_2(B_2) = 1$ and

$$P_{\pi_0}(\pi_t \in B_1) = 1 \quad \text{for all } \pi_0 \in B_1$$
$$P_{\pi_0}(\pi_t \in B_2) = 1 \quad \text{for all } \pi_0 \in B_2.$$

However, by Lemma 2.2.1, this implies that B_1 and B_2 both contain π^*, which is a contradiction. Thus, there must be a unique invariant measure ϕ for $(\pi_t)_{t\geq 0}$.

To prove (ii), first we have that by (i) and the pathwise ergodic theorem (see for example [46, Corollary 2.5.2]) that there exists some $\nu \in \mathcal{P}(\mathcal{X})$ such that for any measurable and bounded function $f : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$, we have

$$\frac{1}{T}\sum_{t=0}^{T-1} f(\pi_t^\nu) \to \int f d\phi$$

P^ν almost surely as $T \to \infty$. But by Lemma 2.2.1 we have that $\tau_\nu \coloneqq \inf\{t \geq 0 : \pi_t^\nu = \pi^*\} < \infty$, and for any prior $\mu \in \mathcal{P}(\mathcal{X})$ we have that and $\tau_\mu \coloneqq \inf\{t \geq 0 : \pi_t^\mu = \pi^*\} < \infty$ almost surely. Thus we have that $\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} f(\pi_t^\nu) = \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} f(\pi_t^\mu)$, and the result follows.

□

Remark 2.2.4. Lemma 2.2.1 also gives some other desirable properties of the process $(\pi_t)_{t\geq 0}$, such as exponential convergence to the invariant measure (see for example [47, Theorem 3.2]), but for our purposes Lemma 2.2.3 will be sufficient.

Recall the definition of $\mathcal{P}_N(\mathcal{X})$ from (2.3), and let $\mathbf{B}_N$ be the set of bins on $\mathcal{P}(\mathcal{X})$ induced by the nearest neighbor map from $\mathcal{P}(\mathcal{X})$ to $\mathcal{P}_N(\mathcal{X})$. Then consider the following set, where ϕ is the unique invariant measure from Lemma 2.2.3,

$$\mathbf{B}_N^{\phi} \coloneqq \{B \in \mathbf{B}_N : \phi(B) > 0\}. \tag{2.6}$$

Also consider the corresponding reproduction values in $\mathcal{P}_N(\mathcal{X})$, given by

$$\mathcal{P}_N^{\phi}(\mathcal{X}) \coloneqq \{\hat{\pi} \in \mathcal{P}_N(\mathcal{X}) : f^{-1}(\hat{\pi}) \in \mathbf{B}_N^{\phi}\}, \tag{2.7}$$

where we have used f to denote the nearest neighbor map from $\mathcal{P}(\mathcal{X})$ to $\mathcal{P}_N(\mathcal{X})$. We now explicitly identify the measures $\nu_{N,i}$ used in (2.4) by defining, for all $B_i \in \mathbf{B}_N^{\phi}$ and $A \in \mathcal{B}(\mathcal{P}(\mathcal{X}))$,

$$\phi_{N,i}(A) \coloneqq \frac{\phi(A)}{\phi(B_i)}.$$

The equations from (2.4) now become

$$\begin{aligned} P_N(\hat{\pi}_j|\hat{\pi}_i, Q) &= \int_{B_i} P(B_j|\pi, Q)\phi_{N,i}(d\pi) \\ c_N(\hat{\pi}_i, Q) &= \int_{B_i} c(\pi, Q)\phi_{N,i}(d\pi). \end{aligned} \tag{2.8}$$

The following will allow us to apply Theorem 1.4.7 to the process $(V_t)_{t\geq 0}$ from (2.5).

Lemma 2.2.5. *For any initialization π_0 in Algorithm 1 and for all $(\hat{\pi}, Q) \in$*

$\mathcal{P}_N^\phi(\mathcal{X}) \times \mathcal{Q}$, the process $(\hat{\pi}_{t+1}, \hat{\pi}_t, Q_t, C_t)_{t\geq 0}$ is such that almost surely,

(i) $(\hat{\pi}_t, Q_t) = (\hat{\pi}, Q)$ infinitely often, and thus $\sum_{t\geq 0} \alpha_t(\hat{\pi}, Q) = \infty$.

(ii)

$$\frac{\sum_{k=0}^{t} C_k \mathbf{1}(\hat{\pi}_k = \hat{\pi}, Q_k = Q)}{\sum_{k=0}^{t} \mathbf{1}(\hat{\pi}_k = \hat{\pi}, Q_k = Q)} \to c_N(\hat{\pi}, Q).$$

(iii)

$$\frac{\sum_{k=0}^{t} f(\hat{\pi}_{k+1}) \mathbf{1}(\hat{\pi}_k = \hat{\pi}, Q_k = Q)}{\sum_{k=0}^{t} \mathbf{1}(\hat{\pi}_k = \hat{\pi}, Q_k = Q)} \to \int_{\mathcal{P}_N^\phi(\mathcal{X})} f(\hat{\pi}_1) P_N(d\hat{\pi}_1 | \hat{\pi}, Q)$$

for any $f : \mathcal{P}_N^\phi(\mathcal{X}) \to \mathbb{R}$.

That is, Assumption 1.4.6 holds for the process $(S_{t+1}, S_t, U_t, C_t)_{t\geq 0} :=$ $(\hat{\pi}_{t+1}, \hat{\pi}_t, Q_t, C_t)_{t\geq 0}$.

Proof. First note that since Q_t is chosen randomly and uniformly from π_t, Lemma 2.2.3 (ii) also implies a similar result for $(\pi_t, Q_t)_{t\geq 0}$, with unique invariant measure given by $\frac{1}{|\mathcal{Q}|}\phi$; that is, for any initialization π_0 and measurable and bounded function $f : \mathcal{P}(\mathcal{X}) \times \mathcal{Q} \to \mathbb{R}$, we have

$$\frac{1}{T} \sum_{t=0}^{T-1} f(\pi_t, Q_t) \to \frac{1}{|\mathcal{Q}|} \int f(\pi, Q) \phi(d\pi)$$

almost surely as $N \to \infty$. By letting f be the respective functions in (i)-(iii), the result follows. □

With Lemma 2.2.5, we are now able to apply Theorem 1.4.7 to obtain the following:

Corollary 2.2.6. *For each* $(\hat{\pi}, Q) \in \mathcal{P}_N^{\phi}(\mathcal{X}) \times \mathcal{Q}$*,* $V_t(\hat{\pi}, Q)$ *defined in* (2.5) *converges almost surely to* $V^*(\hat{\pi}, Q)$ *satisfying,*

$$V^*(\hat{\pi}, Q) = c_N(\hat{\pi}, Q) + \beta \sum_{\hat{\pi}_1 \in \mathcal{P}_N^{\phi}(\mathcal{X})} \min_Q V^*(\hat{\pi}_1, Q) P_N(\hat{\pi}_1 | \hat{\pi}, Q). \tag{2.9}$$

Note that $\min_Q V^*(\hat{\pi}, Q)$ has exactly the form of the DCOE in 1.4.2, however it is only true for each $\hat{\pi} \in \mathcal{P}_N^{\phi}(\mathcal{X})$ (i.e., the bins with ϕ-positive measure). The following result shows that, for certain initializations, it is enough to only consider $\mathcal{P}_N^{\phi}(\mathcal{X})$. Note that the following result holds for *any* $\gamma \in \Gamma_{WS}$, not just the random uniform policy in Algorithm 1. In particular, it will be true for our learned optimal policy.

Lemma 2.2.7. *Let* $\pi_0 = \pi^*$*, where* π^* *is as in Lemma 2.2.1. Under any encoding policy* $\gamma \in \Gamma_{WS}$*, we have that for all* $t \geq 0$*,* $\hat{\pi}_t \in \mathcal{P}_N^{\phi}(\mathcal{X})$ *almost surely.*

Proof. Since all our underlying alphabets are finite, from $\pi_0 = \pi^*$ there are only finitely many possible values of π_1; similarly, there are only finitely many possible values of π_t for each $t \geq 0$. Now, under a uniform selection of Q_t, by Lemma 2.2.1, the return time from each of these (finitely many) values back to π^* is almost surely finite. This implies that the support of ϕ is countable, and that this support is exactly the set of reachable points from π^* under *any* sequence $(Q_t)_{t \geq 0}$.

But this is in particular true for the sequence $(Q_t)_{t \geq 0}$ resulting from some $\gamma \in \Gamma_{WS}$, and thus any π satisfying $P_{\pi^*}^{\gamma}(\pi_t = \pi) > 0$ must be in the support of

ϕ and therefore satisfy $\phi(\pi) > 0$. The result then follows from the definition of $\mathcal{P}_N^\phi(\mathcal{X})$. □

We can now state the main result of this chapter.

Theorem 2.2.8. *Fix any* π_0 *and let* $(\hat{\pi}_t)_{t\geq 0}, (Q_t)_{t\geq 0}$, *and* $(C_t)_{t\geq 0}$ *be generated through Algorithm 1, and let* V_t *be as in* (2.5). *Then the following hold:*

(i) V_t *converges almost surely to a limit* V^*.

(ii) The policy defined by

$$\hat{\gamma}_N^*(\hat{\pi}) \coloneqq \operatorname*{argmin}_{Q\in\mathcal{Q}} V^*(\hat{\pi}, Q) \tag{2.10}$$

is optimal for $MDP_N \coloneqq (\mathcal{P}_N^\phi(\mathcal{X}), \mathcal{Q}, P_N, c_N)$ *for the discounted cost criterion.*

(iii) The policy defined by

$$\widetilde{\gamma}_N^*(\pi) \coloneqq \hat{\gamma}_N^*(\hat{\pi}),$$

where $\hat{\pi}$ *is the nearest neighbor of* π *in* $\mathcal{P}_N(\mathcal{X})$, *satisfies*

$$\lim_{N\to\infty} \left| J_\beta(\pi^*, \widetilde{\gamma}_N^*) - J_\beta^*(\pi^*) \right| = 0,$$

where π^* *is as in Lemma 2.2.1.*

Proof. (i) By Lemmas 2.2.1 and 2.2.7, we have that in finite time $(\hat{\pi}_t)_{t\geq 0}$ will hit π^*, and afterwards will stay within $\mathcal{P}_N^\phi(\mathcal{X})$. Thus outside of $\mathcal{P}_N^\phi(\mathcal{X}) \times \mathcal{Q}$,

$V_t(\hat{\pi}, Q)$ will eventually be constant, and on $\mathcal{P}_N^{\phi}(\mathcal{X}) \times \mathcal{Q}$ convergence follows by Corollary 2.2.6.

(ii) Note that this MDP is restricted to $\mathcal{P}_N^{\phi}(\mathcal{X})$, and this is exactly the set on which we have the DCOE equation (2.9), so we have optimality by Theorem 1.4.2.

(iii) This follows immediately from (ii), Corollary 2.1.2, and the fact that starting at $\pi_0 = \pi^*$, by Lemma 2.2.7, $(\hat{\pi}_t)_{t \geq 0} \subset \mathcal{P}_N^{\phi}(\mathcal{X})$. □

The following is then an immediate corollary of the previous theorem and Theorem 1.4.15.

Corollary 2.2.9. *For every $\epsilon > 0$, there exists some β' such that for all $\beta \in (\beta', 1)$ and all $N \geq N_\beta$,*

$$J(\pi^*, \widetilde{\gamma}^*_{N,\beta}) \leq J^*(\pi^*) + \epsilon,$$

where $\widetilde{\gamma}^*_{N,\beta}$ is the policy from Theorem 2.2.8 (iii) when we compute V_{t+1} using discount parameter β.

Remark 2.2.10. Note that in the average cost case, we can easily modify $\widetilde{\gamma}^*_{N,\beta}$ to be near-optimal for any initial distribution, instead of just $\pi_0 = \pi^*$. Indeed, consider applying a quantizer Q such that $Q^{-1}(m) = \{x\}$, as in the proof of Lemma 2.2.1. In finite time, this will lead to $\pi_t = T(\cdot|x)$, which is a valid value for π^*, and afterwards we apply $\widetilde{\gamma}^*_{N,\beta}$. In the average cost, the suboptimal cost over this finite time disappears, and we obtain near-optimality.

Chapter 3

Belief-Quantization Based Coding for Noisy Channels

As in Chapter 2, we consider an approximation scheme in which the underlying probability space is quantized to some finite set, and present a Q-learning algorithm facilitated by this approximation. We will prove analogous results for the case of a noisy channel. However, some results will be significantly more involved and some statements will be weaker. This is due to the lack of recurrence conditions that were present in the noiseless channel setup which greatly simplified the stochastic analysis both with regard to conditions for Q-learning convergence and implementation for an arbitrary initialization.

3.1 Near-Optimality Results via a Finite MDP Approximation

We recall the definitions of $\mathcal{P}_N(\mathcal{X})$, P_N, and c_N from Section 2.1, which are identical for the noisy channel setup (and where we recall the definitions of P and c from Proposition 1.4.12).

$$\mathcal{P}_N(\mathcal{X}) = \left\{ \hat{\pi} \in \mathcal{P}(\mathcal{X}) : \hat{\pi} = \left[\frac{k_1}{N}, \dots, \frac{k_{|\mathcal{X}|}}{N} \right], k_i = 0, \dots, N, i = 1, \dots, |\mathcal{X}| \right\},$$

$$\begin{aligned} P_N(\hat{\pi}_j | \hat{\pi}_i, Q) &= \int_{B_i} P(B_j | \pi, Q) \nu_{N,i}(d\pi), \\ c_N(\hat{\pi}_i, Q) &= \int_{B_i} c(\pi, Q) \nu_{N,i}(d\pi), \end{aligned}$$

for all $\hat{\pi}_j, \hat{\pi}_i \in \mathcal{P}_N(\mathcal{X})$, where B_j, B_i are the bins of $\hat{\pi}_j, \hat{\pi}_i$, respectively.

The following results hold using the same arguments as Lemma 2.1.1 and Corollary 2.1.2.

Lemma 3.1.1. *The zero-delay coding MDP $(\mathcal{P}(\mathcal{X}), \mathcal{Q}, P, c)$ meets Assumption 1.4.4.*

Corollary 3.1.2. *Let MDP$_N$ $= (\mathcal{P}_N(\mathcal{X}), \mathcal{Q}, P_N, c_N)$. Let $\hat{\gamma}_N^* \in \Gamma_{WS}$ be optimal for MDP$_N$ and let $\widetilde{\gamma}_N^*(\pi) = \hat{\gamma}_N^*(\hat{\pi})$. Then for all $\pi_0 \in \mathcal{P}(\mathcal{X})$ and $\beta \in (0,1)$,*

$$\lim_{N \to \infty} |J_\beta(\pi_0, \widetilde{\gamma}_N^*) - J_\beta^*(\pi_0)| = 0.$$

That is, $\widetilde{\gamma}_N^*$ is near-optimal for the zero-delay coding problem under the dis-

counted distortion criterion.

3.2 Q-learning and its Convergence to a Near-Optimal Finite MDP

Consider the following algorithm to compute the sequences $(\hat{\pi}_t)_{t\geq 0}, (Q_t)_{t\geq 0}$, and $(C_t)_{t\geq 0}$.

Algorithm 2: Quantized Q-learning for noisy channel

```
Require: initial distribution $\pi_0$, transition kernel $T$, channel kernel $O$, quantizer set $\mathcal{Q}$
Sample $X_0 \sim \pi_0$
Choose $Q_0$ uniformly from $\mathcal{Q}$
$M_0 = Q_0(X_0)$
Compute $C_0 = c(\pi_0, Q_0)$ using (1.14)
Sample $M'_0 \sim O(\cdot|M_0)$
for $t \geq 1$ do
    Compute $\pi_t$ using (1.13)
    Sample $X_t \sim T(\cdot|X_{t-1})$
    Choose $Q_t$ uniformly from $\mathcal{Q}$
    $M_t = Q_t(X_t)$
    Compute $C_t = c(\pi_t, Q_t)$ using (1.14)
    Sample $M'_t \sim O(\cdot|M_t)$
```

Then consider the sequence $(V_t)_{t\geq 0}$, where $V_t : \mathcal{P}_N(\mathcal{X}) \times \mathcal{Q} \to \mathbb{R}_+$, defined by

$$V_{t+1}(\hat{\pi}_t, Q_t) = (1 - \alpha_t(\hat{\pi}_t, Q_t))V_t(\hat{\pi}_t, Q_t) + \alpha_t(\hat{\pi}_t, Q_t)\left(C_t + \beta \min_{Q \in \mathcal{Q}} V_t(\hat{\pi}_{t+1}, Q)\right)$$
$$V_{t+1}(\hat{\pi}, Q) = V_t(\hat{\pi}, Q) \quad \text{for all } (\hat{\pi}, Q) \neq (\hat{\pi}_t, Q_t), \tag{3.1}$$

where

$$\alpha_t(\hat{\pi}, Q) = \frac{1}{1 + \sum_{k=0}^{t} \mathbf{1}(\hat{\pi}_k = \hat{\pi}, Q_k = Q)}.$$

Whereas in the noiseless channel case we used the recurrence condition in Lemma 2.2.1 to prove unique invariance (among other results), instead we will use the stability of the predictor process. Recall the discussion of filter and predictor stability in Section 1.5, and recall that $(\pi_t^\mu)_{t\geq 0}$ is used to denote the predictor process with initialization $\pi_0 = \mu$. We will prove the following type of stability:

Definition 3.2.1. We say that the predictor process $(\pi_t)_{t\geq 0}$ is *stable in total variation almost surely* if, for every $\mu, \nu, \kappa \in \mathcal{P}(\mathcal{X})$, we have that P_κ almost surely,

$$\lim_{t\to\infty} \|\pi_t^\mu - \pi_t^\nu\|_{TV} = 0.$$

Note that here P_κ is the measure on $M'_{[0,t-1]}$ induced by $X_0 \sim \kappa$. We make an equivalent definition for the filter process $(\overline{\pi}_t)_{t\geq 0}$.

Theorem 3.2.2. *Under Algorithm 2, the predictor process is stable in total variation almost surely.*

We prove Theorem 3.2.2 with the aid of the following supporting results.

Lemma 3.2.3. *If the filter process is stable in total variation almost surely, then the predictor process is stable in total variation almost surely.*

Proof. Consider the source transition kernel $T(x'|x)$. We have that $\pi^\mu_{t+1}(x') = \sum_x T(x'|x)\overline{\pi}^\mu_t(x)$. By a classical result of Dobrushin [48], this implies that $\|\pi^\mu_{t+1} - \pi^\nu_{t+1}\|_{TV} \leq \|\overline{\pi}^\mu_t - \overline{\pi}^\nu_t\|_{TV}$. The result follows. □

Lemma 3.2.4. *[49, Corollary 5.5] Let $(A_t)_{t\geq 0}$ be a discrete-time Markov chain and $(B_t)_{t\geq 0}$ be a stochastic process such that the B_t are conditionally independent given $(A_t)_{t\geq 0}$. Also assume $P(B_t|A_{[0,\infty)}) = P(B_t|A_t)$, and that $P(B_t|A_t)$ has the form*

$$P(B_t \in B|A_t) = \int_B g(A_t, b)\psi(db),$$

where $g(a,b)$ is a probability density with respect to the σ-finite measure ψ for any a. If g is strictly positive, and $(A_t)_{t\geq 0}$ is aperiodic and Harris recurrent (that is, it visits every state infinitely often with probability one [46, Definition 4.2.10]), then the filter $\overline{\pi}_t(A) := P(A_t \in A|B_{[0,t]})$ is stable in total variation almost surely.

Lemma 3.2.5. *Under Algorithm 2, the filter process $(\overline{\pi}_t)_{t\geq 0}$ is stable in total variation almost surely.*

Proof. We apply Lemma 3.2.4 to $(X_t)_{t\geq 0}$ and $(M'_t)_{t\geq 0}$. Note that under a uniform choice of Q_t, $P(M'_t|X_{[0,\infty)}) = P(M'_t|X_t)$, and we have

$$\begin{aligned}
&P(M'_t = m'|X_t = x) \\
=&\sum_Q P(M'_t = m'|X_t = x, Q_t = Q)P(Q_t = Q|X_t = x) \\
=&\sum_Q O(m'|Q(x))P(Q_t = Q|X_t = x) \\
=&\sum_Q O(m'|Q(x))P(Q_t = Q),
\end{aligned}$$

where the second equality follows from the fact that $M_t = Q_t(X_t)$ is deterministic, and the last equality from the fact that Q_t is chosen independently of X_t in Algorithm 2.

Now, since we are considering the set of all possible quantizers and we choose them all with positive probability, the above expression is always positive (if this were not the case, there is some m' such that $O(m'|m) = 0$ for all $m \in \mathcal{M}$, which implies it is not a valid channel output). This implies the function g in Lemma 3.2.4 is positive. Finally, note that $(X_t)_{t\geq 0}$ evolves independently of the encoding policy; it is always irreducible and aperiodic (thus, since $\mathcal{X}$ is finite, it is Harris recurrent and aperiodic). The result follows from Lemma 3.2.4. □

Lemmas 3.2.3 and 3.2.5 immediately imply Theorem 3.2.2. In the following result, we use predictor stability to prove the uniqueness of an invariant measure for the predictor process.

Theorem 3.2.6. *Under Algorithm 2, $(\pi_t)_{t\geq 0}$ admits a unique invariant measure ϕ.*

Proof. The proof slightly generalizes an argument presented in [50, Corollary 3]. Throughout, we use the notation $\nu(f) := \int f d\nu$. Note that, under Algorithm 2 (where Q_t is chosen randomly), the processes $(\pi_t)_{t\geq 0}$ and $(X_t, \pi_t)_{t\geq 0}$ are Markov, and that as in the noiseless case, by weak continuity and the compactness of $\mathcal{P}(\mathcal{X})$ we know that $(\pi_t)_{t\geq 0}$ has an invariant measure. Thus, we are left with proving uniqueness.

Recall that ζ is the unique invariant measure of our source $(X_t)_{t\geq 0}$. Assume that $m_1, m_2 \in \mathcal{P}(\mathcal{X} \times \mathcal{P}(\mathcal{X}))$ are two invariant measures for $(X_t, \pi_t)_{t\geq 0}$. Then their projections on $\mathcal{X}$ are invariant for $(X_t)_{t\geq 0}$. Then, by unique invariance of ζ we have

$$m_1(dx, d\mu) = P_{m_1}(d\mu|x)\zeta(dx)$$
$$m_2(dx, d\nu) = P_{m_2}(d\nu|x)\zeta(dx)$$

We now show that $m_1(F) = m_2(F)$ for each F on a set of measure-determining functions: $F(x, \nu) = \varphi(x)H(\nu(\varphi_1), \dots, \nu(\varphi_l))$, where $\varphi, \varphi_1, \dots, \varphi_l : \mathcal{X} \to \mathbb{R}$, are continuous and bounded and $H : \mathbb{R}^l \to \mathbb{R}$ is bounded and Lipschitz continuous with constant L_H, and where $l \in \mathbb{Z}_+$ [50].

By invariance we have that

$$m_1(F) = \int F(x_t, \pi_t^\mu) P(dx_t, d\pi_t^\mu | x_0, \mu) P_{m_1}(d\mu|x_0)\zeta(dx_0),$$

and

$$m_2(F) = \int F(x_t', \pi_t^\nu) P(dx_t', d\pi_t^\nu | x_0, \nu) P_{m_2}(d\nu | x_0) \zeta(dx_0),$$

where $P(dx_t, d\pi_t^\mu | x_0, \mu)$ denotes the t-step transition probability given by $P(dx_t, d\pi_t^\mu | x_0, \mu) \coloneqq \int_{(\mathcal{X} \times \mathcal{P}(\mathcal{X}))^{t-1}} P(dx_t, d\pi_t^\mu | x_{t-1}, \pi_{t-1}^\mu) \dots P(dx_1, \pi_1^\mu | x_0, \mu)$, and noting that $\pi_0^\mu = \mu$ by definition. Thus,

$$\begin{aligned}
&|m_1(F) - m_2(F)| \\
\leq &\int |F(x_t, \pi_t^\mu) - F(x_t', \pi_t^\nu)| P(dx_t, d\pi_t^\mu | x_0, \mu) P(dx_t', d\pi_t^\nu | x_0, \nu) \\
&\qquad\qquad\qquad\qquad \cdot P_{m_1}(d\mu | x_0) P_{m_2}(d\nu | x_0) \zeta(dx_0) \\
\leq & L_H \|\varphi\| \int \sum_{i=1}^{l} |\pi_t^\mu(\varphi_i) - \pi_t^\nu(\varphi_i)| P(dx_t, d\pi_t^\mu | x_0, \mu) P(dx_t', d\pi_t^\nu | x_0, \nu) \\
&\qquad\qquad\qquad\qquad \cdot P_{m_1}(d\mu | x_0) P_{m_2}(d\nu | x_0) \zeta(dx_0).
\end{aligned} \tag{3.2}$$

Then note that

$$\begin{aligned}
P(dx_t, d\pi_t^\mu | x_0, \mu) &= \int_{(\mathcal{M}')^t} P(dx_t, d\pi_t^\mu, dm_{[0,t-1]}' | x_0, \mu) \\
&= \int_{(\mathcal{M}')^t} P(dx_t, d\pi_t^\mu | x_0, \mu, m_{[0,t-1]}') P(dm_{[0,t-1]}' | x_0, \mu) \\
&= \int_{(\mathcal{M}')^t} P(dx_t, d\pi_t^\mu | x_0, \mu, m_{[0,t-1]}') P(dm_{[0,t-1]}' | x_0),
\end{aligned}$$

where the last line follows from the fact that given $X_0 = x_0$ (and under a random choice of $Q_{[0,t-1]}$), the joint measure on $M_{[0,t-1]}'$ no longer depends on the initial distribution μ. Thus, we can rewrite (3.2) as

$$L_H\|\varphi\| \int \sum_{i=1}^{l} |\pi_t^{\mu}(\varphi_i) - \pi_t^{\nu}(\varphi_i)| P(dx_t, d\pi_t^{\mu}|x_0, \mu, m'_{[0,t-1]})$$
$$\cdot\, P(dx'_t, d\pi_t^{\nu}|x_0, \nu, m'_{[0,t-1]}) P(dm'_{[0,t-1]}|x_0) P_{m_1}(d\mu|x_0) P_{m_2}(d\nu|x_0)\zeta(dx_0)$$

Note that now the measures on π_t^{μ} and π_t^{ν} are conditioned on the same $m'_{[0,t-1]}$, so we can invoke Theorem 3.2.2 to claim that the integrand,

$$\int |\pi_t^{\mu}(\varphi_i) - \pi_t^{\nu}(\varphi_i)| P(dx_t, d\pi_t^{\mu}|x_0, \mu, m'_{[0,t-1]}) P(dx'_t, d\pi_t^{\nu}|x_0, \nu, m'_{[0,t-1]})$$
$$\cdot\, P(dm'_{[0,t-1]}|x_0) P_{m_1}(d\mu|x_0) P_{m_2}(d\nu|x_0),$$

goes to zero as $t \to \infty$ for every x_0 and for every $i = 1, \ldots, l$. Thus, by the dominated convergence theorem, we have that (3.2) goes to zero as $t \to \infty$, and we have that m_1 and m_2 are in fact the same measure.

Next we show that $(\pi_t)_{t\geq 0}$ admits at most one invariant measure. Assume that $v_1, v_2 \in \mathcal{P}(\mathcal{P}(\mathcal{X}))$ are two different invariant measures for $(\pi_t)_{t\geq 0}$. Then there exists a continuous and bounded $f : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ such that $v_1(f) \neq v_2(f)$. Now for $j = 1, 2$, let $(X_t^j, \pi_t^j)_{t\geq 0}$ be the process with initial law $\pi(dx)v_j(d\pi)$. Since $\mathcal{X}$ is finite, $P(X_t^j \in \cdot\,, \pi_t^j \in \cdot\,)$ is tight.

Now, since $\mathcal{X}$ is finite, we also have that $(X_t, \pi_t)_{t\geq 0}$ has a weakly continuous transition kernel. Thus the time average

$$\frac{1}{T}\sum_{t=0}^{T-1} P(X_t^j \in \cdot\,, \pi_t^j \in \cdot\,)$$

converges weakly to an invariant measure η_j for $(X_t, \pi_t)_{t\geq 0}$ [33, Theorem 3.3.1].

Then for $F(x,\pi) = f(\pi)$, we have $\eta_1(F) = v_1(f) \neq v_2(f) = \eta_2(F)$. But then η_1 and η_2 are two different invariant measures for $(X_t, \pi_t)_{t\geq 0}$, which is a contradiction. Thus $(\pi_t)_{t\geq 0}$ admits at most one invariant measure. □

Unlike in the noiseless channel case, we cannot explicitly identify an element of ϕ-positive measure. However, the following lemma shows that ζ, the invariant distribution of the source, is in the support of ϕ. Thus ζ will play a similar role to the π^* of Lemma 2.2.1.

Lemma 3.2.7. *We have $\zeta \in$ supp(ϕ), where supp(ϕ) denotes the support of ϕ in $\mathcal{P}(\mathcal{X})$.*

Proof. Consider some open neighborhood of radius δ containing ζ, say $N_\delta(\zeta) \subset \mathcal{P}(\mathcal{X})$. Now consider a totally "uninformative" quantizer $Q \in \mathcal{Q}$; that is $Q(x) = i$ for all $x \in \mathcal{X}$ and some $i \in \mathcal{M}$. Via the update equation (4), if $Q_t = Q$, we have that $\pi_{t+1} = \pi_t P$, where we have used the matrix notation $P(i,j) \coloneqq P(X_{t+1} = j | X_t = i)$. Since $(X_t)_{t\geq 0}$ is irreducible and aperiodic, πP^T converges in total variation to ζ as $T \to \infty$, and so there exists some $T > 0$ such that for all $\pi \in \mathcal{P}(\mathcal{X})$, $\pi P^T \in N_\delta(\zeta)$. Thus, if $Q_t = Q$ for

$t = 0, \dots, T-1$, then we have $\pi_T \in N_\delta(\zeta)$.

But under Algorithm 2, we have some fixed positive probability of choosing Q, say $P(Q_t = Q) = p > 0$. In particular, for all $t \geq T$, $P(\pi_t \in N_\delta(\zeta)|\pi_0 = \pi) \geq P(Q_{[t-T,t-1]} = (Q, Q, \dots, Q)) = p^T$. This implies that ζ is "accessible" (since any neighborhood of this element will be visited with positive probability from any initial prior, see [51, Definition 2.1]) and hence that $\zeta \in \operatorname{supp}(\phi)$ [51, Lemma 2.2]. $\square$

For technical reasons, we require the following assumption on the unique invariant measure ϕ. Recall that $\mathbf{B}_N$ is the set of bins under the nearest neighbor map from $\mathcal{P}(\mathcal{X})$ to $\mathcal{P}_N(\mathcal{X})$.

Assumption 3.2.8. *For all $B \in \mathbf{B}_N$, we have $\phi(\partial B) = 0$, where ∂B denotes the boundary of B.*

Remark 3.2.9. Assumption 3.2.8 is in general not easy to verify directly. However, note that if Assumption 3.2.8 does *not* hold, we have $\phi(\partial B) > 0$, while $\lambda(\partial B) = 0$. Further, we have that the update equation (1.13) can be rewritten as some matrix equation $\pi_{t+1} = \pi_t \mathsf{M}_{Q,m'}$, where $\mathsf{M}_{Q,m'}$ is the update matrix corresponding to $Q_t = Q$ and $M'_t = m'$. If we could show that $\operatorname{rank}(\mathsf{M}_{Q,m'}) \geq |\mathcal{X}| - 1$ for all Q and m', this would imply that the preimage of ∂B, say $B_{Q,m'} := \mathsf{M}^{-1}_{Q,m'}(\partial B)$, also satisfies $\phi(B_{Q,m'}) > 0$ and $\lambda(B_{Q,m'}) = 0$. This implies that the support of ϕ, say $\operatorname{supp}(\phi)$, would also satisfy $\lambda(\operatorname{supp}(\phi)) = 0$. Therefore, if instead of the bin boundaries induced by $\mathcal{P}_N(\mathcal{X})$, one uniformly chose the bin boundaries from the respective spaces, then Assumption 3.2.8 would hold almost surely. We leave this as an interesting future research direction.

Lemma 3.2.10. *Let Assumption 3.2.8 hold. Then under Algorithm 2 and for any belief quantization bin B, as $T \to \infty$,*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbf{1}_B(\pi_t^{\zeta}) \to \phi(B)\ P_{\zeta}\text{-a.s..}$$

That is, starting from $\pi_0 = \zeta$, the empirical occupation measures of the predictor process will converge to its invariant distribution on the belief quantization bins.

Proof. By the pathwise ergodic theorem (see [46, Corollary 2.5.2]), for ϕ-almost every $\mu \in \mathcal{P}(\mathcal{X})$, we have that as $N \to \infty$,

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbf{1}_B(\pi_t^{\mu}) \to \phi(B)\ P_{\mu}\text{-a.s.} \tag{3.3}$$

Note that the above also holds for ψ-almost every μ for any $\psi \ll \phi$. Accordingly, define ψ_k to be the restriction of ϕ to the open ball of radius $\frac{1}{k}$ centered at ζ:

$$\psi_k(A) = \frac{\phi(A)}{\phi(N_k(\zeta))}$$

where N_k is the open ball of radius $\frac{1}{k}$ around ζ and for any measurable $A \subset N_k(\zeta)$. Note that $\phi(N_k(\zeta)) > 0$ for all $k \geq 1$ by Lemma 3.2.7. Now, by definition $\psi_k \ll \phi$ for all $k \geq 1$. Therefore, for all $k \geq 1$, there exists at least one $\mu_k \in N_k(\zeta)$ such that (3.3) holds with $\mu = \mu_k$. This implies the existence of a sequence $(\mu_k)_{k\geq 0}$ such that (3.3) holds for all k and such that $||\mu_k - \zeta||_{TV} \to 0$.

Since the transition kernel $P(d\pi'|\pi, Q)$ is weakly continuous (Lemma 1.4.14), so is $P(d\pi'|\pi)$ when Q_t is chosen uniformly. Thus the empirical occupation measures given by $\nu_k^{(n)}(A) := \frac{1}{N}\sum_{t=0}^{N-1} \mathbf{1}_A(\pi_t^{\mu_k}) \to \frac{1}{N}\sum_{t=0}^{N-1} \mathbf{1}_A(\pi_t^{\zeta})$ weakly as $k \to \infty$. By the Portmanteau theorem [32, Theorem 1.4.16], we have that weak convergence implies convergence on any set $A \subset \mathcal{P}(\mathcal{X})$ with $\phi(\partial A) = 0$. Under Assumption 3.2.8, this holds in particular for every belief quantization bin B, and the result follows. □

Recall (2.6) and (2.7) from the noiseless channel case, which are defined identically for the noisy channel case:

$$\mathbf{B}_N^{\phi} := \{B \in \mathbf{B}_N : \phi(B) > 0\}$$

and

$$\mathcal{P}_N^{\phi}(\mathcal{X}) := \{\hat{\pi} \in \mathcal{P}_N(\mathcal{X}) : f^{-1}(\hat{\pi}) \in \mathbf{B}_N^{\phi}\}.$$

We similarly define $\phi_{N,i}$, P_N, and c_N as:

$$\begin{aligned}
\phi_{N,i}(A) &:= \frac{\phi(A)}{\phi(B_i)} \\
P_N(\hat{\pi}_j|\hat{\pi}_i, Q) &= \int_{B_i} P(B_j|\pi, Q)\phi_{N,i}(d\pi) \\
c_N(\hat{\pi}_i, Q) &= \int_{B_i} c(\pi, Q)\phi_{N,i}(d\pi).
\end{aligned}$$

The following is then the analog of Lemma 2.2.5 from the noiseless channel case. Note that here we enforce that $\pi_0 = \zeta$, whereas in Lemma 2.2.5 π_0 could be arbitrary.

Lemma 3.2.11. *Let $\pi_0 = \zeta$ in Algorithm 2 and let Assumption 3.2.8 hold. Then for all $(\hat{\pi}, Q) \in \mathcal{P}_N^{\phi}(\mathcal{X}) \times \mathcal{Q}$, the process $(\hat{\pi}_{t+1}, \hat{\pi}_t, Q_t, C_t)_{t\geq 0}$ is such that almost surely,*

(i) $(\hat{\pi}_t, Q_t) = (\hat{\pi}, Q)$ *infinitely often, and thus* $\sum_{t\geq 0} \alpha_t(\hat{\pi}, Q) = \infty$.

(ii)

$$\frac{\sum_{k=0}^{t} C_k \mathbf{1}(\hat{\pi}_k = \hat{\pi}, Q_k = Q)}{\sum_{k=0}^{t} \mathbf{1}(\hat{\pi}_k = \hat{\pi}, Q_k = Q)} \to c_N(\hat{\pi}, Q).$$

(iii)

$$\frac{\sum_{k=0}^{t} f(\hat{\pi}_{k+1}) \mathbf{1}(\hat{\pi}_k = \hat{\pi}, Q_k = Q)}{\sum_{k=0}^{t} \mathbf{1}(\hat{\pi}_k = \hat{\pi}, Q_k = Q)} \to \int_{\mathcal{P}_N^{\phi}(\mathcal{X})} f(\hat{\pi}_1) P_N(d\hat{\pi}_1 | \hat{\pi}, Q)$$

for any $f : \mathcal{P}_N^{\phi}(\mathcal{X}) \to \mathbb{R}$.

That is, Assumption 1.4.6 holds for the process $(S_{t+1}, S_t, U_t, C_t)_{t\geq 0} \coloneqq (\hat{\pi}_{t+1}, \hat{\pi}_t, Q_t, C_t)_{t\geq 0}$.

Proof. The proof follows identically to Lemma 2.2.5, but invoking Lemma 3.2.10 (and requiring Assumption 3.2.8) instead of Lemma 2.2.3. □

We can then apply Theorem 1.4.7 to obtain the noisy-channel version of Corollary 2.2.6.

Corollary 3.2.12. *For each $(\hat{\pi}, Q) \in \mathcal{P}_N^{\phi}(\mathcal{X}) \times \mathcal{Q}$, $V_t(\hat{\pi}, Q)$ defined in* (3.1) *converges almost surely to $V^*(\hat{\pi}, Q)$ satisfying,*

$$V^*(\hat{\pi}, Q) = c_N(\hat{\pi}, Q) + \beta \sum_{\mathcal{P}_N^{\phi}(\mathcal{X})} \min_{Q} V^*(\hat{\pi}_1, Q) P_N(\hat{\pi}_1 | \hat{\pi}, Q). \tag{3.4}$$

We now present a result similar to Lemma 2.2.7 in the sense that it allows us to restrict our state space to $\mathcal{P}_N^{\phi}(\mathcal{X})$, but the result is slightly weaker and the proof is more involved.

Lemma 3.2.13. *Under* any $\gamma \in \Gamma_{WS}$*, for any $\pi \in \mathcal{P}(\mathcal{X})$ which is reachable from $\mu \in supp(\phi)$ (that is, any π with $P^{\gamma}(\pi_t = \pi | \pi_0 = \mu) > 0$), we have either*

(i) $\pi \in B$ for some $B \in \mathbf{B}_N^{\phi}$, or

(ii) π is on the boundary of some $B \in \mathbf{B}_N^{\phi}$.

Proof. Let $\{N_k\}_{k \geq 0} \subset \mathcal{P}(\mathcal{X})$ be a sequence of open balls centered at μ such that $\bigcap_{k=0}^{\infty} N_k = \mu$ and let $\{\psi_k\}_{k \geq 0}$ be defined as $\psi_k(A) = \frac{\phi(A)}{\phi(N_k)}$ for all $A \subset N_k$; that is, ψ_k is the restriction of ϕ to N_k. Note that by definition, $\phi(N_k) > 0$ and $\psi_k \ll \phi$ for all k. We also have by weak continuity of $P(d\pi' | \pi, Q)$ that $P_{\psi_k}(d\pi_t | Q_{[0,t-1]} = \overline{Q})$ converges weakly to $P_{\mu}(d\pi_t | Q_{[0,t-1]} = \overline{Q})$, where we have used P_{μ} to denote the conditional measure on π_t induced by $\pi_0 = \mu$ and P_{ψ_k} to denote (with an abuse of notation) the measure induced by $\pi_0 \sim \psi_k$.

That is,

$$P_{\psi_k}(d\pi_t|Q_{[0,t-1]}=\overline{Q}) = \int \psi_k(d\pi)P(d\pi_t|\pi_0=\pi, Q_{[0,t-1]}=\overline{Q})$$
$$P_{\mu}(d\pi_t|Q_{[0,t-1]}=\overline{Q}) = P(d\pi_t|\pi_0=\mu, Q_{[0,t-1]}=\overline{Q})$$

Now let $B \subset \mathcal{P}(\mathcal{X})$ be open. By the Portmanteau theorem (e.g., [46, Theorem 1.4.16]), we have that

$$\liminf_{k\to\infty} P_{\psi_k}(\pi_t \in B|Q_{[0,t-1]}=\overline{Q}) \geq P_{\mu}(\pi_t \in B|Q_{[0,t-1]}=\overline{Q}). \tag{3.5}$$

Now by invariance of ϕ, we have that for any $T \geq 1$,

$$\phi(B) = \frac{1}{T}\sum_{t=0}^{T-1}\sum_{\overline{Q}\in\mathcal{Q}^t}\frac{1}{|\mathcal{Q}|^t}\int \phi(d\pi)P(\pi_t \in B|\pi_0=\pi, Q_{[0,t-1]}=\overline{Q}), \tag{3.6}$$

where the $\frac{1}{|\mathcal{Q}|^t}$ is due to marginalizing over all $Q_{[0,t-1]}$ and since we choose Q_t uniformly from $\mathcal{Q}$ at each $t \geq 0$. But we also have for each ψ_k,

$$P_{\psi_k}(\pi_t \in B|Q_{[0,t-1]}=\overline{Q}) = \int \psi_k(d\pi)P(\pi_t \in B|\pi_0=\pi, Q_{[0,t-1]}=\overline{Q}). \tag{3.7}$$

Thus we have the following chain of implications for all sufficiently large k,

$$P_{\mu}(\pi_t \in B|Q_{[0,t-1]}=\overline{Q}) > 0 \Rightarrow P_{\psi_k}(\pi_t \in B|Q_{[0,t-1]}=\overline{Q}) > 0 \Rightarrow \phi(B) > 0$$

where the first follows from (3.5), and the second from (3.6), (3.7), and the fact that $\psi_k \ll \phi$ for all $k \geq 0$.

But since $\overline{Q}$ was arbitrary, this also holds for any policy $\gamma \in \Gamma_{WS}$. Thus,

$$P_\mu^\gamma(\pi_t \in B) > 0 \Rightarrow \phi(B) > 0.$$

Finally, let π be reachable under γ, i.e., $P_\mu^\gamma(\pi_t = \pi) > 0$. Then the above implies that, for any open neighborhood $N(\pi)$ around π, we have $\phi(N(\pi)) > 0$. This implies that π must satisfy either (i) or (ii). □

Corollary 3.2.14. *Let $\pi_0 = \mu \in supp(\phi)$ and $\gamma \in \Gamma_{WS}$, and consider the resulting process $(\pi_t^\mu)_{t\geq 0}$. Let $(\hat{\pi}_t^\mu)_{t\geq 0}$ be the corresponding nearest neighbors in $\mathcal{P}_N(\mathcal{X})$. Then with an appropriate choice of tie-breaking rules on the nearest neighbor map, we have that $(\hat{\pi}_t^\mu)_{t\geq 0} \subset \mathcal{P}_N^\phi(\mathcal{X})$.*

Proof. The result follows immediately from Lemma 3.2.13 by recognizing that, with an appropriate choice of tie-breaking rule, we can enforce that only case (i) occurs. □

For the remainder, we assume that such a tie-breaking rule is used, and provide a simple method for obtaining one such rule at the end of this chapter. The following is then the analog of Theorem 2.2.8.

Theorem 3.2.15. *Let $\pi_0 = \zeta$ and let Assumption 3.2.8 hold. Let $(\hat{\pi}_t)_{t\geq 0}$, $(Q_t)_{t\geq 0}$, and $(C_t)_{t\geq 0}$ be generated through Algorithm 2, and let V_t be as in* (3.1)*. Then the following hold:*

(i) V_t converges almost surely to a limit V^*.

(ii) The policy defined by

$$\hat{\gamma}_N^*(\hat{\pi}) \coloneqq \operatorname*{argmin}_{Q \in \mathcal{Q}} V^*(\hat{\pi}, Q) \tag{3.8}$$

is optimal for $MDP_N \coloneqq (\mathcal{P}_N^{\phi}(\mathcal{X}), \mathcal{Q}, P_N, c_N)$ *for the discounted cost criterion.*

(iii) The policy defined by

$$\widetilde{\gamma}_N^*(\pi) \coloneqq \hat{\gamma}_N^*(\hat{\pi}),$$

where $\hat{\pi}$ *is the nearest neighbor of* π *in* $\mathcal{P}_N(\mathcal{X})$*, satisfies*

$$\lim_{N \to \infty} \left| J_\beta(\zeta, \widetilde{\gamma}_N^*) - J_\beta^*(\zeta) \right| = 0.$$

Proof. The results follow from the same arguments in the proof of Theorem 2.2.8, by invoking the relevant theorems from this section. In particular, (i) follows from Corollary 3.2.12, (ii) follows from (3.4) and the DCOE, and (iii) follows from part (ii), Corollary 3.1.2, and Lemmas 3.2.7 and 3.2.13. □

The following is then an immediate corollary of the previous theorem and Theorem 1.4.15.

Corollary 3.2.16. *For every* $\epsilon > 0$*, there exists some* β' *such that for all* $\beta \in (\beta', 1)$ *and all* $N \geq N_\beta$*,*

$$J(\zeta, \widetilde{\gamma}_{N,\beta}^*) \leq J^*(\zeta) + \epsilon,$$

where $\widetilde{\gamma}^*_{N,\beta}$ is the policy from Theorem 3.2.15 (iii) when we compute V_{t+1} using discount parameter β.

Remark 3.2.17. To determine the tie-breaking rule in Corollary 3.2.14, note that by Lemma 3.2.10 we have that with probability one, the average empirical occupation times of each $\hat{\pi} \in \mathcal{P}^{\phi}_{N}(\mathcal{X})$ will converge to a positive value, while those of $\hat{\pi} \in \mathcal{P}_{N}(\mathcal{X}) \setminus \mathcal{P}^{\phi}_{N}(\mathcal{X})$ will converge to zero. Accordingly, for the tie-breaking rule in Corollary 3.2.14, one can simply break ties by choosing the bin with a greater empirical occupation time in order to ensure we always stay within $\mathcal{P}^{\phi}_{N}(\mathcal{X})$.

Chapter 4

Sliding Finite Window Scheme

4.1 Approximation

Under this scheme we approximate π_t using a sliding finite window, rather than a nearest neighbor scheme. The analysis in this section is inspired by [44], which used a similar construction to study sliding finite window policies for partially observed Markov decision processes (POMDPs).

4.1.1 An Alternative Exact MDP and its Optimality

First, we must define a slightly different MDP than the previous chapters, where we studied $(\mathcal{P}(\mathcal{X}), \mathcal{Q}, P, c)$. Fix some window size $N \in \mathbb{Z}_+$. Recall the channel outputs $(M'_t)_{t\geq 0}$ and the quantizers $(Q_t)_{t\geq 0}$. We define the following:

$$I_t \coloneqq (M'_{[t-N,t-1]}, Q_{[t-N,t-1]})$$

$$W_t \coloneqq (\pi_{t-N}, I_t).$$

Note that we can compute π_t given W_t by applying the update equations in (1.13) N times. Denote this mapping by

$$\varphi : \mathcal{W} \to \mathcal{P}(\mathcal{X})$$
$$W_t \mapsto \pi_t$$

where $\mathcal{W} = \mathcal{P}(\mathcal{X}) \times (\mathcal{M}')^N \times \mathcal{Q}^N$, endowed with the product topology, where we use the weak convergence topology on $\mathcal{P}(\mathcal{X})$ and standard coordinate topologies on $\mathcal{M}'$ and $\mathcal{Q}$.

We call W_t the sliding finite window belief term, and call policies of the form $Q_t = \gamma_t(W_t)$ *sliding finite window belief* policies (with window size N). If it does not depend on t, we call it stationary. Denote the set of all stationary sliding finite window belief policies by Γ_{FS}.

Remark 4.1.1. This approach assumes that we start at time $t \geq N$; although for a general MDP the first N steps may be significant, for the zero-delay coding problem we are interested in taking $\beta \to 1$, so these first N steps will not be crucial. Accordingly, we assume that N steps have already been completed with some arbitrary $\gamma \in \Gamma_{WS}$. For notational simplicity, we assume that these steps have occurred from $t = -N, \ldots, -1$ and thus the process starts from W_0 (and the prior would now be π_{-N}).

This sliding finite window belief construction inherits the MDP properties of the original setup. Indeed, it is straightforward to show that the process $(W_t)_{t\geq 0}$ is controlled Markov, with control $(Q_t)_{t\geq 0}$. That is, for all $t \geq 0$,

$$P(W_{t+1}|W_{[0,t]}, Q_{[0,t]}) = P(W_{t+1}|W_t, Q_t). \tag{4.1}$$

Proposition 4.1.2. *Under any $\gamma \in \Gamma_{FS}$, the zero-delay coding problem is an MDP, where:*

1. *$\mathcal{Z} = \mathcal{W}$.*
2. *$\mathcal{U} = \mathcal{Q}$.*
3. *$P = P(dw'|w, Q)$.*
4. *$c(w, Q) = \sum_{\mathcal{M}'} \min_{\hat{x}\in\hat{\mathcal{X}}} \sum_{\mathcal{X}} d(x, \hat{x}) O_Q(m'|x)\varphi(w)(x)$,*

where we recall the notation $O_Q(m'|x) = O(m'|Q(x))$.

Note that the cost function is exactly the cost function we had in Proposition 1.4.12, by simply replacing $\pi = \varphi(w)$. Then an analog to Lemma 1.4.13 holds, and we indeed have that solving the MDP from Proposition 4.1.2 is equivalent to solving the zero-delay coding problem. That is, we can equivalently consider $J^*_\beta(w_0) = \inf_{\gamma\in\Gamma_{FS}} J_\beta(w_0, \gamma)$.

The next proposition follows immediately from Proposition 1.4.10 and the fact that $\pi_t = \varphi(W_t)$.

Proposition 4.1.3. *For any $\beta \in (0, 1)$ and $N \in \mathbb{Z}_+$, there exists $\gamma^* \in \Gamma_{FS}$*

that solves the discounted cost problem; that is, one that satisfies

$$J_\beta(w_0, \gamma^*) = J^*_\beta(w_0)$$

for all $w_0 \in \mathcal{W}$*.*

4.1.2 Near-Optimal Finite Model (Sliding Window) Approximation of the Alternative MDP

The above representation is still not particularly useful, as it still requires one to compute π_{t-N}. Instead, fix the first coordinate to ζ and let

$$\hat{W}_t = (\zeta, I_t) \tag{4.2}$$

$$\hat{\pi}_t = \varphi(\hat{W}_t). \tag{4.3}$$

That is, we obtain $\hat{\pi}_t$ by applying the update equations N times, but starting from an incorrect (fixed) prior ζ. Equivalently,

$$\pi_t(x) = P^\gamma_{\pi_{t-N}}(X_t = x | M'_{[t-N,t-1]}, Q_{[t-N,t-1]})$$
$$\hat{\pi}_t(x) = P^\gamma_\zeta(X_t = x | M'_{[t-N,t-1]}, Q_{[t-N,t-1]}).$$

The key idea, which will be discussed in detail later, is that under predictor stability the correct predictor $\pi_t = \varphi(W_t)$ and the incorrect predictor $\hat{\pi}_t = \varphi(\hat{W}_t)$ will be close for large enough N, since the predictor will be insensitive to the prior.

The benefits of such an approximation are evident: rather than deal with all of $\mathcal{W}$, which is uncountable due to $\mathcal{P}(\mathcal{X})$, we only have to deal with the finite set $\mathcal{W}_N := \{\zeta\} \times (\mathcal{M}')^N \times \mathcal{Q}^N$. Furthermore, we no longer need to compute π_{t-N}, which can save significant computation resources especially when the relevant alphabets are large.

Consider the following transition kernel,

$$P_N(\hat{w}_1|\hat{w}, Q) := P(\mathcal{P}(\mathcal{X}), i_1|\hat{w}, Q), \tag{4.4}$$

where P is the transition kernel of the sliding finite window belief MDP and $\hat{w}_1 = (\zeta, i_1)$, and cost function

$$c_N(\hat{w}, Q) := \sum_{\mathcal{M}'} \min_{\hat{x} \in \hat{\mathcal{X}}} \sum_{\mathcal{X}} d(x, \hat{x}) O_Q(m'|x) \hat{\pi}(x). \tag{4.5}$$

Then, our approximate MDP becomes $\text{MDP}_N = (\mathcal{W}_N, \mathcal{Q}, P_N, c_N)$. Denote the discounted cost for this MDP under a given policy (from $\mathcal{W}_N$ to $\mathcal{Q}$) by $\hat{J}_\beta(\hat{w}_0, \hat{\gamma})$, and the optimal discounted cost by $\hat{J}^*_\beta(\hat{w}_0)$, with minimizing policy $\hat{\gamma}^*_N$. The relevant extensions (which are obtained by making the previous functions constant over $\mathcal{P}(\mathcal{X})$) are then $\widetilde{J}^*_\beta(w_0)$ and $\widetilde{\gamma}^*_N$. The following is a key loss term:

$$L_t^N := \sup_{\gamma \in \Gamma_{WS}} \mathbf{E}^{\gamma}_{\pi_{t-N}} \left[||\pi_t - \hat{\pi}_t||_{TV} \right]. \tag{4.6}$$

We now present our main results for this approximation scheme, which give a bound on the performance loss when using the given window length N. Note that here we take an expectation with respect to some policy acting on the previous N steps (and hence generates W_0), and with respect to some prior π_{-N}. Also, define $||d||_\infty := \max_{x,\hat{x}} d(x,\hat{x})$, where we recall that d is the distortion measure for the zero-delay coding problem.

Theorem 4.1.4. *For any $\gamma \in \Gamma_{WS}$ acting on N time steps to generate W_0 and any prior $\pi_{-N} \in \mathcal{P}(\mathcal{X})$, we have*

$$\mathbf{E}^{\gamma}_{\pi_{-N}}\left[\left|\widetilde{J}^*_\beta(W_0) - J^*_\beta(W_0)\right|\right] \leq \frac{||d||_\infty}{1-\beta}\sum_{t=0}^{\infty}\beta^t L^N_t.$$

Theorem 4.1.5. *For any $\gamma \in \Gamma_{WS}$ acting on N time steps to generate W_0 and any prior $\pi_{-N} \in \mathcal{P}(\mathcal{X})$, we have*

$$\mathbf{E}^{\gamma}_{\pi_{-N}}\left[\left|J_\beta(W_0,\widetilde{\gamma}^*_N) - J^*_\beta(W_0)\right|\right] \leq \frac{2||d||_\infty}{1-\beta}\sum_{t=0}^{\infty}\beta^t L^N_t.$$

The proofs for Theorems 4.1.4 and 4.1.5 are given at the end of this chapter.

4.1.3 Bounds on the Loss Term

The loss term L^N_t in the previous theorems is related to the question of predictor stability (recall Section 1.5). Indeed, the term within the supremum is exactly

$$\mathbf{E}_{\mu}^{\gamma}\left[||\pi_t^{\mu} - \pi_t^{\nu}||_{TV}\right] \tag{4.7}$$

when $\mu = \pi_{t-N}$ and $\nu = \zeta$, and under some $\gamma \in \Gamma_{WS}$. Thus bounding this term over all γ will give us a bound on L_t^N. We note that any notion of predictor stability could be used to give a bound on L_t^N (and thus on the performance of $\widetilde{\gamma}_N^*$). In the following section, we give one such condition which is sufficient (but by no means necessary) to apply the above theorems.

Dobrushin Coefficient Conditions

The following results are inspired by the analysis in [52], which uses joint contraction properties of the state and observation kernels to bound (4.7). First we introduce some notation. For standard Borel spaces $\mathcal{A}_1, \mathcal{A}_2$ and some kernel $K : \mathcal{A}_1 \to \mathcal{P}(\mathcal{A}_2)$, we define the Dobrushin coefficient as

$$\delta(K) := \inf \sum_{i=1}^{n} \min(K(B_i|x), K(B_i|y)),$$

where the infimum is over $x, y \in \mathcal{A}_1$ and all partitions $\{B_i\}_{i=1}^n$ of $\mathcal{A}_2$. In particular, for finite spaces, the Dobrushin coefficient is equivalent to summing the minimum elements between every pair of rows, then taking the minimum of these sums. For example, take

$$K = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}.$$

Between the first and second rows, the sum of the minimum elements gives $\frac{2}{3}$, between the third and fourth gives $\frac{3}{4}$, etc. One can verify that the minimum of such sums is $\frac{1}{2}$, so $\delta(K) = \frac{1}{2}$ (note that $\delta(K) \leq 1$ by definition).

The following is then a counterpart of [52, Theorem 3.6]. Note that in our case, the channel is not time-invariant, unlike in [52], but the analysis follows similarly. The proof is provided at the end of the chapter.

Theorem 4.1.6. *For any $\gamma \in \Gamma_{WS}$ and for any $\mu \ll \nu$, we have*

$$\mathbf{E}_{\mu}^{\gamma}\left[||\pi_{t+1}^{\mu} - \pi_{t+1}^{\nu}||_{TV}\right] \leq (1 - \delta(T))(2 - \tilde{\delta}(O))\mathbf{E}_{\mu}^{\gamma}\left[||\pi_{t}^{\mu} - \pi_{t}^{\nu}||_{TV}\right],$$

where $\tilde{\delta}(O) = \min_{Q \in \mathcal{Q}}(\delta(O_Q))$.

We can arrive at a simpler bound by using $\delta(O)$ directly rather than $\tilde{\delta}(O)$. To see this, note that for a given quantizer Q, the kernel $O_Q(m'|x) = O(m'|Q(x))$ only contains rows from the kernel O, thus $\delta(O) \leq \delta(O_Q)$ for all Q. Thus we arrive at the following corollary.

Corollary 4.1.7. *Assume $\alpha := (1 - \delta(T))(2 - \delta(O)) < 1$. Then for any*

$\gamma \in \Gamma_{WS}$ and for any $\mu \ll \nu$, we have

$$\mathbf{E}_{\mu}^{\gamma}\left[||\pi_{t+1}^{\mu} - \pi_{t+1}^{\nu}||_{TV}\right] \leq \alpha \mathbf{E}_{\mu}^{\gamma}\left[||\pi_{t}^{\mu} - \pi_{t}^{\nu}||_{TV}\right].$$

That is, the predictor process is exponentially stable in total variation in expectation. Furthermore, if $\delta(T) > \frac{1}{2}$, then the above is true with $\alpha = 1 - \delta(T)$ regardless of the channel O.

Applying this to the L_t^N term, we have

$$\begin{aligned} L_t^N &= \sup_{\gamma \in \Gamma_{WS}} \mathbf{E}_{\pi_{t-N}}^{\gamma}\left[||\pi_t - \hat{\pi}_t||_{TV}\right] \\ &\leq \alpha^N ||\pi_{t-N} - \zeta||_{TV} \leq 2\alpha^N. \end{aligned} \tag{4.8}$$

Note that, in many applications of the zero-delay quantization problem, the requirement that $(1 - \delta(T))(2 - \delta(O)) < 1$ is too strong. In particular, in the special case where the channel is noiseless, we will always have that $\delta(O) = 0$. Thus we can only use Corollary 4.1.7 if $\delta(T) > \frac{1}{2}$.

This is not surprising given the nature of Dobrushin-type conditions. At a high level, the Dobrushin coefficient tells us how similar the conditional measures $O_Q(m'|x)$ and $O_Q(m'|y)$ are for different $x, y \in \mathcal{X}$. The more similar they are, the closer the coefficient is to 1. Therefore, such Dobrushin-type conditions prioritize *uninformative* kernels, as these will have Dobrushin coefficients closer to 1. Conversely, the goal of the zero-delay coding problem is to use quantizers that create *informative* kernels. Nevertheless, the above conditions give an easy-to-verify condition for predictor stability, and give us a rather strong form of stability (exponential).

Combining Theorem 4.1.5 and the bound in (4.8), we obtain the following result.

Corollary 4.1.8. *Assume $\alpha := (1-\delta(T))(2-\delta(O)) < 1$. Then for any $\gamma \in \Gamma_{WS}$ which acts on N time steps to generate w_0 and any prior π_{-N}, we have*

$$\mathbf{E}^{\gamma}_{\pi_{-N}}\left[\left|J_\beta(W_0,\widetilde{\gamma}^*_N) - J^*_\beta(W_0)\right|\right] \le \frac{4||d||_\infty}{(1-\beta)^2}\alpha^N.$$

4.2 Q-Learning and its Convergence to the Sliding Finite Window Model MDP

As in previous chapters, we present an algorithm to compute three sequences. However, in this algorithm, our three sequences are $(\hat{W}_t)_{t\geq 0}$, $(Q_t)_{t\geq 0}$, and $(C_t)_{t\geq 0}$. Note also the time shift - we assume that the algorithm starts at time $t = -N$.

Algorithm 3: Sliding finite window Q-learning

```
Require: initial distribution $\pi_{-N}$, transition kernel $T$, channel kernel $O$,
         quantizer set $\mathcal{Q}$
Sample $X_{-N} \sim \pi_{-N}$
Choose $Q_{-N}$ uniformly from $\mathcal{Q}$
```

```
$M_{-N} = Q_{-N}(X_{-N})$
Sample $M'_{-N} \sim O(\cdot|M_{-N})$
for $t = -N+1, \ldots, -1$ do
    Sample $X_t \sim T(\cdot|X_{t-1})$
    Choose $Q_t$ uniformly from $\mathcal{Q}$
    $M_t = Q_t(X_t)$
    Sample $M'_t \sim O(\cdot|M_t)$
for $t \geq 0$ do
    Compute $\hat{W}_t$ using (4.2)
    Sample $X_t \sim T(\cdot|X_{t-1})$
    Choose $Q_t$ uniformly from $\mathcal{Q}$
    $M_t = Q_t(X_t)$
    Compute $C_t = c_N(\hat{W}_t, Q_t)$ using (4.5)
    Sample $M'_t \sim O(\cdot|M_t)$
```

Then consider the sequence $(V_t)_{t \geq 0}$, where $V_t : \mathcal{W}_N \times \mathcal{Q} \to \mathbb{R}_+$, defined by

$$\begin{aligned} V_{t+1}(\hat{W}_t, Q_t) &= (1 - \alpha_t(\hat{W}_t, Q_t))V_t(\hat{W}_t, Q_t) \\ &\qquad + \alpha_t(\hat{W}_t, Q_t)\left(C_t + \beta \min_{Q \in \mathcal{Q}} V_t(\hat{W}_{t+1}, Q)\right) \end{aligned} \tag{4.9}$$

$$V_{t+1}(\hat{w}, Q) = V_t(\hat{w}, Q) \quad \text{for all } (\hat{w}, Q) \neq (\hat{W}_t, Q_t), \tag{4.10}$$

where

$$\alpha_t(\hat{w}, Q) = \frac{1}{1 + \sum_{k=0}^{t} \mathbf{1}(\hat{W}_k = \hat{w}, Q_k = Q)}.$$

Remark 4.2.1. For the nearest neighbor scheme, we used the "true" cost $c(\pi_t, Q_t)$, since computing the approximate cost c_N from (2.3) is difficult for two reasons: (i) the map from $\mathcal{P}(\mathcal{X})$ to $\mathcal{P}_N(\mathcal{X})$ makes computing the corresponding bins somewhat complicated, and (ii) more importantly, we do not know the measures $\phi_{N,i}$ in (2.4), only that they exist. For the sliding finite window scheme, we wish to avoid computing π_t (since it is not needed for the approximation), and computing c_N is straightforward via (4.5). Accordingly, we use the "approximate" cost $c_N(\hat{w}_t, Q_t)$. Further advantages/disadvantages of this approach will be discussed in the following chapter.

Note that not every element of $\mathcal{W}_N$ has positive probability of occurring - some sequences of channel outputs and quantizers may be impossible (for any prior) depending on our source and channel. Note that having zero probability under any prior π_{-N} is equivalent to having zero probability under ζ, since $\pi \ll \zeta$ for all π. Accordingly, we define the following set, which will play a similar role to $\mathcal{P}_N^{\phi}(\mathcal{X})$ in Chapters 2 and 3.

$$\mathcal{W}_N^+ := \{\hat{w} \in \mathcal{W}_N : P_\zeta(\hat{W}_0 = \hat{w}) > 0\}.$$

Theorem 4.2.2. *Under Algorithm 3, for any π_{-N} and for all $(\hat{w}, Q) \in \mathcal{W}_N^+ \times \mathcal{Q}$, the process $(\hat{W}_{t+1}, \hat{W}_t, Q_t, C_t)_{t\geq 0}$ is such that almost surely,*

(i) $(\hat{W}_t, Q_t) = (\hat{w}, Q)$ infinitely often, and thus $\sum_{t\geq 0} \alpha_t(\hat{w}, Q) = \infty$.

(ii)

$$\frac{\sum_{k=0}^{t} C_k \mathbf{1}(\hat{W}_k = \hat{w}, Q_k = Q)}{\sum_{k=0}^{t} \mathbf{1}(\hat{W}_k = \hat{w}, Q_k = Q)} \to c_N(\hat{w}, Q).$$

(iii)

$$\frac{\sum_{k=0}^{t} f(\hat{W}_{k+1}) \mathbf{1}(\hat{W}_k = \hat{w}, Q_k = Q)}{\sum_{k=0}^{t} \mathbf{1}(\hat{W}_k = \hat{w}, Q_k = Q)} \to \int_{\mathcal{W}_N^+} f(\hat{w}_1) P_N(d\hat{w}_1 | \hat{w}, Q)$$

for any $f : \mathcal{W}_N^+ \to \mathbb{R}$.

Proof. To show (i), note that since the source $(X_t)_{t \geq 0}$ is positive Harris recurrent (with invariant measure ζ), the marginals on X_t are such that $||P(X_t \in \cdot) - \zeta||_{TV} \to 0$ as $t \to \infty$. Accordingly, every $\hat{w} \in \mathcal{W}_N^+$ will eventually satisfy $P_{\pi_{t-N}}(\hat{W}_t = \hat{w}) > 0$ for all sufficiently large t, and since there are only finitely many this implies that each is hit infinitely often by $\hat{W}_t$. We have (ii) directly by $C_k = c_N(\hat{W}_k, Q_k)$.

(iii) follows from a similar argument to (i); by noting that the marginals on X_t converge to ζ, we must have for any $f : (\mathcal{M}')^N \times \mathcal{Q}^N \to \mathbb{R}$,

$$\frac{\sum_{k=0}^{t} f(I_{k+1}) \mathbf{1}(I_k = i, Q_k = Q)}{\sum_{k=0}^{t} \mathbf{1}(I_k = i, Q_k = Q)} \to \int f(i_1) P(\mathcal{P}(\mathcal{X}), i_1 | \zeta, i, Q),$$

where $P(\mathcal{P}(\mathcal{X}), i_1 | \zeta, i, Q)$ is the kernel from (4.1). But this is exactly the definition of P_N, so the result follows. □

We can then apply Theorem 1.4.7 to obtain an analogous result to Corollaries 2.2.6 and 3.2.12,

Corollary 4.2.3. *For each* $(\hat{w}, Q) \in \mathcal{W}_N^+ \times \mathcal{Q}$*,* $V_t(\hat{w}, Q)$ *defined in* (4.10) *converges almost surely to* $V^*(\hat{w}, Q)$ *satisfying,*

$$V^*(\hat{w}, Q) = c_N(\hat{w}, Q) + \beta \sum_{\mathcal{W}_N^+} \min_{Q} V^*(\hat{w}_1, Q) P_N(\hat{w}_1 | \hat{w}, Q). \tag{4.11}$$

Thus we obtain the following analogous result to Theorems 2.2.8 and 3.2.15, and the proof follows identically.

Theorem 4.2.4. *Fix any* π_{-N} *and let* $(\hat{W}_t)_{t\geq 0}, (Q_t)_{t\geq 0}$*, and* $(C_t)_{t\geq 0}$ *be generated through Algorithm 3, and let* V_t *be as in* (4.10)*. Further, assume that the source and channel kernels satisfy* $\alpha := (1 - \delta(T))(2 - \delta(O)) < 1$*. Then the following hold:*

(i) V_t *converges almost surely to a limit* V^**.*

(ii) The policy defined by

$$\hat{\gamma}_N^*(\hat{w}) := \operatorname*{argmin}_{Q \in \mathcal{Q}} V^*(\hat{w}, Q) \tag{4.12}$$

is optimal for $MDP_N := (\mathcal{W}_N^+, \mathcal{Q}, P_N, c_N)$ *for the discounted cost criterion.*

(iii) The policy defined by

$$\widetilde{\gamma}_N^*(w) := \hat{\gamma}_N^*(\hat{w}),$$

where $\hat{w} = (\zeta, i)$ and $w = (\pi, i)$ agree on their i coordinates, satisfies

$$\mathbf{E}^{\gamma}_{\pi_{-N}}\left[\left|J_\beta(W_0, \widetilde{\gamma}^*_N) - J^*_\beta(W_0)\right|\right] \leq \frac{4||d||_\infty}{(1-\beta)^2}\alpha^N,$$

and thus for each $w_0 \in \mathcal{W}^+_N$,

$$\lim_{N\to\infty}\left|J_\beta(w_0, \widetilde{\gamma}^*_N) - J^*_\beta(w_0)\right| = 0.$$

The following is then an immediate corollary of the previous theorem and Theorem 1.4.15.

Corollary 4.2.5. *For every $\epsilon > 0$, there exists some β' such that for all $\beta \in (\beta', 1)$, all $N \geq N_\beta$, and all $w_0 \in \mathcal{W}^+_N$,*

$$J(w_0, \widetilde{\gamma}^*_{N,\beta}) \leq J^*(w_0) + \epsilon,$$

where $\widetilde{\gamma}^*_{N,\beta}$ is the policy from Theorem 4.2.4 (iii) when we compute V_{t+1} using discount parameter β.

4.3 Proofs of Theorems 4.1.4 and 4.1.5

Lemma 4.3.1. *Recall $I_t = (M'_{[t-N,t-1]}, Q_{[t-N,t-1]})$, $W_t = (\pi_{t-N}, I_t)$, $\hat{W}_t = (\zeta, I_t)$, and $\hat{\pi}_t = P^\gamma_\zeta(X_t|m'_{[t-N,t-1]}, Q_{[t-N.t-1]})$. Then for any $w_t \in \mathcal{W}$, $\hat{w}_t \in$*

$\mathcal{W}_N$*, and* $Q_t \in \mathcal{Q}$ *we have*

$$||P(M_t' \in \cdot|w_t, Q_t) - P(M_t' \in \cdot|\hat{w}_t, Q_t)||_{TV} \leq ||\pi_t - \hat{\pi}_t||_{TV}.$$

Proof. Let $f : \mathcal{M}' \to \mathbb{R}$ be measurable with $||f||_\infty \leq 1$. Then,

$$\begin{aligned}
&\left|\sum_{\mathcal{M}'} f(m_t')P(m_t'|w_t, Q_t) - \sum_{\mathcal{M}'} f(m_t')P(m_t'|\hat{w}_t, Q_t)\right| \\
&= \left|\sum_{\mathcal{X}}\sum_{\mathcal{M}'} f(m_t')P(m_t'|w_t, x_t, Q_t)P(x_t|w_t, Q_t)\right. \\
&\qquad\qquad\qquad \left. - \sum_{\mathcal{X}}\sum_{\mathcal{M}'} f(m_t')P(m_t'|\hat{w}_t, x_t, Q_t)P(x_t|\hat{w}_t, Q_t)\right| \\
&= \left|\sum_{\mathcal{X}}\sum_{\mathcal{M}'} f(m_t')O_{Q_t}(m_t'|x_t)\pi_t(x_t) - \sum_{\mathcal{X}}\sum_{\mathcal{M}'} f(m_t')O_{Q_t}(m_t'|x_t)\hat{\pi}_t(x_t)\right| \\
&\leq ||\pi_t - \hat{\pi}_t||_{TV},
\end{aligned}$$

where the third line follows from conditional independence of M_t' and W_t given (X_t, Q_t), and since Q_t is a function of W_t under any $\gamma \in \Gamma_{WS}$. The last line follows from the fact that $g(x) := \sum_{\mathcal{M}'} f(m')O_Q(m'|x)$ is upper bounded by 1. Taking the supremum over all such f yields the result. □

Proof of Theorem 4.1.4

We provide a proof for the case when $N = 1$, but an analogous proof follows for $N > 1$. It can be shown (see [32, Theorem 4.2.3]) that the functions

$J_\beta(w_t,\gamma)$ and $J^*_\beta(w_t)$ satisfy the following fixed-point equations:

$$\begin{aligned} J_\beta(w_t,\gamma) &= c(w_t,\gamma(w_t)) + \beta \int_{\mathcal{W}} J_\beta(w_{t+1},\gamma) P(dw_{t+1}|w_t,\gamma(w_t)) \\ J^*_\beta(w_t) &= \min_{Q_t\in\mathcal{Q}}\Big(c(w_t,Q_t) + \beta \int_{\mathcal{W}} J^*_\beta(w_{t+1}) P(dw_{t+1}|w_t,Q_t)\Big), \end{aligned}$$

for all $w_t \in \mathcal{W}$ and $\gamma \in \Gamma_{FS}$. Note that although the integral is over $\mathcal{W}$, which is uncountable, we can only reach finitely many elements from a given w_t since the observation space $\mathcal{M}'$ is finite. In particular, when $N = 1$, we can write $w_t = (\pi_{t-1}, m'_{t-1}, Q_{t-1})$ and $w_{t+1} = (\pi_t, m'_t, Q_t)$, so the above becomes

$$J_\beta(w_t,\gamma) = c(w_t,\gamma(w_t)) + \beta \sum_{m'_t\in\mathcal{M}'} J_\beta((\pi_t, m'_t, \gamma(w_t)),\gamma) P(m'_t|w_t,\gamma(w_t)) \tag{4.13}$$

$$J^*_\beta(w_t) = \min_{Q_t\in\mathcal{Q}}\Big(c(w_t,Q_t) + \beta \sum_{m'_t\in\mathcal{M}'} J^*_\beta(\pi_t, m'_t, Q_t) P(m'_t|w_t,Q_t)\Big). \tag{4.14}$$

The functions $\hat{J}_\beta(\hat{w}_t,\hat{\gamma})$ and $\hat{J}^*_\beta(\hat{w}_t)$ satisfy equivalent fixed-point equations to (4.14), so that for $N = 1$,

$$\hat{J}_\beta(\hat{w}_t,\hat{\gamma}) = c_1(\hat{w}_t,\hat{\gamma}(\hat{w}_t)) + \beta \sum_{m'_t\in\mathcal{M}'} \hat{J}_\beta(\zeta, m'_t, \hat{\gamma}(\hat{w}_t)) P(m'_t|\hat{w}_t,\hat{\gamma}(\hat{w}_t))$$

$$\hat{J}^*_\beta(\hat{w}_t) = \min_{Q_t\in\mathcal{Q}}\Big(c_1(\hat{w}_t,Q_t) + \beta \sum_{m'_t\in\mathcal{M}'} \hat{J}^*_\beta(\zeta, m'_t, Q_t) P(m'_t|\hat{w}_t,Q_t)\Big). \tag{4.15}$$

By definition of the extension $\widetilde{J}^*_\beta$ we have $\hat{J}^*_\beta(\hat{w}_1) = \widetilde{J}^*_\beta(w_1)$. Thus,

$$\beta \sum_{m_0' \in \mathcal{M}'} \hat{J}^*_\beta(\zeta, m_0', Q_0) P(m_0'|w_0, Q_0) = \beta \sum_{m_0' \in \mathcal{M}'} \widetilde{J}^*_\beta(\pi_0, m_0', Q_0) P(m_0'|w_0, Q_0).$$

We add and subtract the above term and use $\widetilde{J}^*_\beta(w_0) = \hat{J}^*_\beta(\hat{w}_0)$ to obtain

$$\begin{aligned}
&\left|\widetilde{J}^*_\beta(w_0) - J^*_\beta(w_0)\right| \\
&= \Big|\hat{J}^*_\beta(\hat{w}_0) - \beta \sum_{m_0' \in \mathcal{M}'} \hat{J}^*_\beta(\zeta, m_0', Q_0) P(m_0'|w_0, Q_0) \\
&\qquad + \beta \sum_{m_0' \in \mathcal{M}'} \widetilde{J}^*_\beta(\pi_0, m_0', Q_0) P(m_0'|w_0, Q_0) - J^*_\beta(w_0)\Big|.
\end{aligned}$$

Then applying the fixed-point equations (4.14) and (4.15) to the last and first terms respectively,

$$\begin{aligned}
&\left|\widetilde{J}^*_\beta(w_0) - J^*_\beta(w_0)\right| \\
&\le \max_{Q_0 \in \mathcal{Q}} |c_1(\hat{w}_0, Q_0) - c(w_0, Q_0)| \\
&+ \beta \max_{Q_0 \in \mathcal{Q}} \left|\sum_{m_0'} \hat{J}^*_\beta(\zeta, m_0', Q_0) P(m_0'|\hat{w}_0, Q_0) - \sum_{m_0'} \hat{J}^*_\beta(\zeta, m_0', Q_0) P(m_0'|w_0, Q_0)\right| \\
&+ \beta \max_{Q_0 \in \mathcal{Q}} \left|\sum_{m_0'} \widetilde{J}^*_\beta(\pi_0, m_0', Q_0) P(m_0'|w_0, Q_0) - \sum_{m_0'} J^*_\beta(\pi_0, m_0', Q_0) P(m_0'|w_0, Q_0)\right|.
\end{aligned}$$

We now bound these three terms in expectation. The expectation is on W_0 and $\hat{W}_0$, with respect to the prior π_{-1} and some $\gamma \in \Gamma_{WS}$, but we omit these in the expectation for notational simplicity. For the first term, by definition of c and c_N we have

$$\mathbf{E}\left[\left|c_1(\hat{W}_0, Q_0) - c(W_0, Q_0)\right|\right] \leq ||d||_\infty \mathbf{E}\left[||\hat{\pi}_0 - \pi_0||_{TV}\right] \leq ||d||_\infty L_0^1, \quad (4.16)$$

where $||d||_\infty = \max_{x,\hat{x}} d(x, \hat{x})$ and we recall the definition of L_t^N in (4.6); that is, $L_0^1 = \sup_{\gamma \in \Gamma_{WS}} \mathbf{E}\left[||\pi_t - \hat{\pi}_t||_{TV}\right]$. For the second term, we have

$$\begin{aligned}
&\mathbf{E}\left[\max_{Q_0 \in \mathcal{Q}} \left|\sum_{m_0'} \hat{J}_\beta^*(\zeta, m_0', Q_0) P(m_0'|\hat{W}_0, Q_0) - \sum_{m_0'} \hat{J}_\beta^*(\zeta, m_0', Q_0) P(m_0'|W_0, Q_0)\right|\right] \\
&\leq ||\hat{J}_\beta^*||_\infty \mathbf{E}\left[\max_{Q_0 \in \mathcal{Q}} ||P(m_0'|\hat{W}_0, Q_0) - P(m_0'|W_0, Q_0)||_{TV}\right] \\
&\leq ||\hat{J}_\beta^*||_\infty L_0^1,
\end{aligned}$$

where the first inequality follows from the definition of the total variation (since $\hat{J}_\beta^* / ||\hat{J}_\beta^*||_\infty$ is bounded by 1) and the second inequality is due to Lemma 4.3.1. Finally, since both sums in the last term are over the same measure $P(m_0'|W_0, Q_0)$, we have

$$\begin{aligned}
&\mathbf{E}\left[\max_{Q_0 \in \mathcal{Q}} \left|\sum_{m_0'} \widetilde{J}_\beta^*(\pi_0, m_0', Q_0) P(m_0'|W_0, Q_0) - \sum_{m_0'} J_\beta^*(\pi_0, m_0', Q_0) P(m_0'|W_0, Q_0)\right|\right] \\
&\leq \sup_{\gamma' \in \Gamma_{WS}} \mathbf{E}\left[\left|\widetilde{J}_\beta^*(W_1) - J_\beta^*(W_1)\right|\right],
\end{aligned}$$

where we have used $(\pi_0, M_0', Q_0) = W_1$. Combining all three bounds (and multiplying by β where appropriate) gives us

$$\begin{aligned}&\mathbf{E}\left[\left|\widetilde{J}^*_\beta(W_0) - J^*_\beta(W_0)\right|\right] \\ \leq&(||d||_\infty + \beta||\hat{J}^*_\beta||_\infty)L^1_0 + \beta \sup_{\gamma' \in \Gamma_{WS}} \mathbf{E}\left[\left|\widetilde{J}^*_\beta(W_1) - J^*_\beta(W_1)\right|\right]\end{aligned}$$

We apply the same process to the final term, then recursively and by the fact that $||J^*_\beta||_\infty \leq \frac{||d||_\infty}{1-\beta}$, we have

$$\mathbf{E}\left[\left|\widetilde{J}^*_\beta(w_0) - J^*_\beta(W_0)\right|\right] \leq \frac{||d||_\infty}{1-\beta}\sum_{t=0}^{\infty}\beta^t L^1_t.$$

Proof of Theorem 4.1.5

As before, we consider $N = 1$, but analogous arguments follow for $N > 1$. We apply a similar strategy, by using the fixed-point equations in the proof of Theorem 2. Also, let $Q^*_0 \coloneqq \widetilde{\gamma}^*_1(w_0)$; that is, the action given by making the optimal policy for MDP$_1$ constant over $\mathcal{P}(\mathcal{X})$. Then, by (4.13),

$$J_\beta(w_0, \widetilde{\gamma}^*_1) = c(w_0, Q^*_0) + \beta \sum_{m'_0 \in \mathcal{M}'} J_\beta((\pi_0, m'_0, Q^*_0), \widetilde{\gamma}^*_1)P(m'_0|w_0, Q^*_0)$$

and using (4.15) and the fact that $\widetilde{J}^*_\beta(w_0) = \hat{J}^*_\beta(\hat{w}_0)$,

$$\widetilde{J}^*_\beta(w_0) = c_1(\hat{w}_0, Q^*_0) + \beta \sum_{m'_0 \in \mathcal{M}'} \widetilde{J}^*_\beta(\pi_0, m'_0, Q^*_0)P(m'_0|\hat{w}_0, Q^*_0).$$

Using $w_1 = (\pi_0, m_0', Q_0^*)$, we add and subtract

$$\sum_{m_0' \in \mathcal{M}'} \widetilde{J}_\beta^*(w_1) P(m_0' | w_0, Q_0^*),$$

to obtain

$$\begin{aligned}
&\left| J_\beta(w_0, \widetilde{\gamma}_1^*) - \widetilde{J}_\beta^*(w_0) \right| \\
\leq\ & |c(w_0, Q_0^*) - c(\hat{w}_0, Q_0^*)| \\
&+ \beta \sum_{m_0' \in \mathcal{M}'} \left| \widetilde{J}_\beta^*(w_1) P(m_0' | \hat{w}_0, Q_0^*) - \widetilde{J}_\beta^*(w_1) P(m_0' | w_0, Q_0^*) \right| \\
&+ \beta \sum_{m_0' \in \mathcal{M}'} \left| \widetilde{J}_\beta^*(w_1) - J_\beta(w_1, \widetilde{\gamma}_1^*) \right| P(m_0' | w_0, Q_0^*).
\end{aligned}$$

Thus, using (4.16) and Lemma 4.3.1,

$$\begin{aligned}
&\mathbf{E}\left[\left| J_\beta(W_0, \widetilde{\gamma}_1^*) - \widetilde{J}_\beta^*(W_0) \right| \right] \\
\leq\ & ||d||_\infty L_0^1 \\
&+ \beta ||\widetilde{J}_\beta^*||_\infty \sup_{\gamma \in \Gamma_{WS}} \mathbf{E}\left[||P(m_0' | \hat{W}_0, Q_0^*) - P(m_0' | W_0, Q_0^*)||_{TV} \right] \\
&+ \beta \sup_{\gamma \in \Gamma_{WS}} \mathbf{E}\left[\left| \widetilde{J}_\beta^*(W_1) - J_\beta(W_1, \widetilde{\gamma}_1^*) \right| \right] \\
\leq & (||d||_\infty + \beta ||\widetilde{J}_\beta^*||_\infty) L_0^1 + \beta \sup_{\gamma'} \mathbf{E}\left[\left| J_\beta(W_1, \widetilde{\gamma}_1^*) - \widetilde{J}_\beta^*(W_1) \right| \right].
\end{aligned}$$

Recursively, and using the fact that $||\widetilde{J}_{\beta}^{*}||_{\infty} \leq \frac{||d||_{\infty}}{1-\beta}$,

$$\mathbf{E}\left[\left|J_{\beta}^{*}(W_0, \widetilde{\gamma}_1^{*}) - J_{\beta}^{*}(W_0)\right|\right] \leq \frac{||d||_{\infty}}{1-\beta} \sum_{t=0}^{\infty} \beta^t L_t^1. \tag{4.17}$$

Finally, we have

$$\begin{aligned} &\mathbf{E}\left[\left|J_{\beta}^{*}(W_0, \widetilde{\gamma}_1^{*}) - J_{\beta}^{*}(W_0)\right|\right] \\ \leq\, &\mathbf{E}\left[\left|J_{\beta}^{*}(W_0, \widetilde{\gamma}_1^{*}) - \widetilde{J}_{\beta}^{*}(W_0)\right|\right] + \mathbf{E}\left[\left|\widetilde{J}_{\beta}^{*}(W_0) - J_{\beta}^{*}(W_0)\right|\right] \\ \leq\, &\frac{2||d||_{\infty}}{1-\beta} \sum_{t=0}^{\infty} \beta^t L_t^1, \end{aligned}$$

where the final inequality follows from (4.17) and Theorem 4.1.4. ■

4.4 Proof of Theorem 4.1.6

Lemma 4.4.1. *The following holds:*

$$\sum_{\mathcal{X}} \sum_{\mathcal{M}'} ||\overline{\pi}_t^{\mu} - \overline{\pi}_t^{\nu}||_{TV} O_{Q_t}(m'|x) \pi_t^{\mu}(x) \leq (2 - \tilde{\delta}(O)) ||\pi_t^{\mu} - \pi_t^{\nu}||_{TV},$$

where $\tilde{\delta}(O) = \min_{Q \in \mathcal{Q}}(\delta(O_Q))$.

Proof. The following argument is from [52, Lemma 3.5], adapted to our setup. Let $f : \mathcal{X} \to \mathbb{R}$ be measurable with $||f||_{\infty} \leq 1$. Recall the update equations for $\pi_t, \overline{\pi_t}$ given in (1.13), and let $N^{\mu}(M_t, Q_t)$ denote the normalizing term for the $(\pi_t^{\mu})_{t \geq 0}$ process, $N^{\mu}(M_t', Q_t) := \sum_{\mathcal{X}} g_{Q_t}(x, M_t') \pi_t^{\mu}(x)$. Then we have for any $M_t' = m'$ and $Q_t = Q$,

$$\begin{aligned}
&\left|\sum_{\mathcal{X}} f(x)\overline{\pi}_t^{\mu}(x) - \sum_{\mathcal{X}} f(x)\overline{\pi}_t^{\nu}(x)\right| \\
=&\left|\sum_{\mathcal{X}} f(x)\frac{g_Q(x,m')\pi_t^{\mu}(x)}{N^{\mu}(m',Q)} - \sum_{\mathcal{X}} f(x)\frac{g_Q(x,m')\pi_t^{\nu}(x)}{N^{\nu}(m',Q)}\right| \\
\leq &\left|\sum_{\mathcal{X}} f(x)\frac{g_Q(x,m')\pi_t^{\mu}(x)}{N^{\mu}(m',Q)} - \sum_{\mathcal{X}} f(x)\frac{g_Q(x,m')\pi_t^{\nu}(x)}{N^{\mu}(m',Q)}\right| \\
&\qquad + \left|\sum_{\mathcal{X}} f(x)\frac{g_Q(x,m')\pi_t^{\nu}(x)}{N^{\mu}(m',Q)} - \sum_{\mathcal{X}} f(x)\frac{g_Q(x,m')\pi_t^{\nu}(x)}{N^{\nu}(m',Q)}\right| \\
=&\frac{1}{N^{\mu}(m',Q)}\left|\sum_{\mathcal{X}} f(x)g_Q(x,m')\pi_t^{\mu}(x) - \sum_{\mathcal{X}} f(x)g_Q(x,m')\pi_t^{\nu}(x)\right| \\
&\qquad + \left|\frac{1}{N^{\mu}(m',Q)} - \frac{1}{N^{\nu}(m',Q)}\right| \cdot \left|\sum_{\mathcal{X}} f(x)g_Q(x,m')\pi_t^{\nu}(x)\right| \\
\leq&\frac{1}{N^{\mu}(m',Q)}\sum_{\mathcal{X}} g_Q(x,m')|\pi_t^{\mu} - \pi_t^{\nu}|(x) \\
&\qquad + \left|\frac{1}{N^{\mu}(m',Q)} - \frac{1}{N^{\nu}(m',Q)}\right| N^{\nu}(m',Q) \\
\leq&\frac{1}{N^{\mu}(m',Q)}\left(\sum_{\mathcal{X}} g_Q(x,m')|\pi_t^{\mu} - \pi_t^{\nu}|(x) + |N^{\mu}(m',Q) - N^{\nu}(m',Q)|\right),
\end{aligned}$$

where in the second last line we have used the notation $\sum_{\mathcal{X}} |\pi_t^{\mu} - \pi_t^{\nu}|(x) = \sum_{\mathcal{X}}(\mathbf{1}_{S^+} - \mathbf{1}_{S^-})(\pi_t^{\mu} - \pi_t^{\nu})(x)$ with $S^+ = \{x|(\pi_t^{\mu} - \pi_t^{\nu})(x) > 0\}$ and $S^- = \{x|(\pi_t^{\mu} - \pi_t^{\nu})(x) \leq 0\}$. Note that $\sum_{\mathcal{X}} |\pi_t^{\mu} - \pi_t^{\nu}|(x) = ||\pi_t^{\mu} - \pi_t^{\nu}||_{TV}$. Taking the supremum over all f gives

$$||\overline{\pi}_t^{\mu} - \overline{\pi}_t^{\nu}||_{TV} \leq \frac{1}{N^{\mu}(m',Q)}\Bigg(\sum_{\mathcal{X}} g_Q(x,m')|\pi_t^{\mu} - \pi_t^{\nu}|(x)$$

$$+\left|N^{\mu}(m',Q)-N^{\nu}(m',Q)\right|\Bigg). \quad (4.18)$$

Thus, recalling that ψ is some appropriate reference measure over $\mathcal{M}'$,

$$\begin{aligned}
&\sum_{\mathcal{X}}\sum_{\mathcal{M}'}||\overline{\pi}^{\mu}_t-\overline{\pi}^{\nu}_t||_{TV}O_{Q_t}(m'|x)\pi^{\mu}_t(x)\\
=&\sum_{\mathcal{X}}\sum_{\mathcal{M}'}||\overline{\pi}^{\mu}_t-\overline{\pi}^{\nu}_t||_{TV}g_{Q_t}(x,m')\psi(m')\pi^{\mu}_t(x)\\
=&\sum_{\mathcal{M}'}||\overline{\pi}^{\mu}_t-\overline{\pi}^{\nu}_t||_{TV}\left(\sum_{\mathcal{X}}g_{Q_t}(x,m')\pi^{\mu}_t(x)\right)\psi(m')\\
=&\sum_{\mathcal{M}'}||\overline{\pi}^{\mu}_t-\overline{\pi}^{\nu}_t||_{TV}N^{\mu}(m',Q_t)\psi(m')\\
\leq&\sum_{\mathcal{M}'}\Big(\sum_{\mathcal{X}}g_{Q_t}(x,m')|\pi^{\mu}_t-\pi^{\nu}_t|(x)+|N^{\mu}(m',Q_t)-N^{\nu}(m',Q_t)|\Big)\psi(m')\\
\leq&\sum_{\mathcal{X}}\left(\sum_{\mathcal{M}'}g_{Q_t}(x,m')\psi(m')\right)|\pi^{\mu}_t-\pi^{\nu}_t|(x)\\
&\qquad\qquad+\sum_{\mathcal{M}'}\left|\sum_{\mathcal{X}}g_{Q_t}(x,m')(\pi^{\mu}_t-\pi^{\nu}_t)(x)\right|\psi(m')\\
\leq&||\pi^{\mu}_t-\pi^{\nu}_t||_{TV}+\sum_{\mathcal{M}'}\left|O_{Q_t}(\pi^{\mu}_t)-O_{Q_t}(\pi^{\nu}_t)\right|(m')\\
=&||\pi^{\mu}_t-\pi^{\nu}_t||_{TV}+||O_{Q_t}(\pi^{\mu}_t)-O_{Q_t}(\pi^{\nu}_t)||_{TV},
\end{aligned}$$

where in the second last line we have used the kernel operator notation $O_Q(\pi)(m')=\sum_{\mathcal{X}}O_Q(m'|x)\pi(x)$. It is shown in [48] that the Dobrushin coefficient acts as a contraction coefficient for kernel operators under total variation. In particular

$$||O_{Q_t}(\pi_t^\mu) - O_{Q_t}(\pi_t^\nu)||_{TV} \leq (1 - \delta(O_{Q_t}))||\pi_t^\mu - \pi_t^\nu||_{TV}. \tag{4.19}$$

Thus,

$$\begin{aligned}
&\sum_{\mathcal{X}} \sum_{\mathcal{M}'} ||\overline{\pi}_t^\mu - \overline{\pi}_t^\nu||_{TV} O_{Q_t}(m'|x)\pi_t^\mu(x) \\
\leq &(2 - \delta(O_{Q_t}))||\pi_t^\mu - \pi_t^\nu||_{TV} \\
\leq &(2 - \tilde{\delta}(O))||\pi_t^\mu - \pi_t^\nu||_{TV},
\end{aligned}$$

where $\tilde{\delta}(O) = \min_{Q \in \mathcal{Q}}(\delta(O_Q))$. □

Proof of Theorem 4.1.6

Note that, in $\mathbf{E}_\mu^\gamma$ expectations of $\overline{\pi}_t^\mu$ and $\overline{\pi}_t^\nu$, it is enough to take the expectation over only $M'_{[0,t]}$, since under any $\gamma \in \Gamma_{WS}$, $Q_{[0,t]}$ are deterministic given μ and $M'_{[0,t-1]}$. Thus,

$$\begin{aligned}
&\mathbf{E}_\mu^\gamma \left[||\overline{\pi}_t^\mu - \overline{\pi}_t^\mu||_{TV}\right] \\
= &\sum_{(\mathcal{M}')^{t+1}} ||\overline{\pi}_t^\mu - \overline{\pi}_t^\mu||_{TV} P_\mu^\gamma(m'_{[0,t]}) \\
= &\sum_{(\mathcal{M}')^t} \sum_{\mathcal{X}} \sum_{\mathcal{M}'} ||\overline{\pi}_t^\mu - \overline{\pi}_t^\mu||_{TV} P_\mu^\gamma(m'_t|x_t, m'_{[0,t-1]}) P_\mu^\gamma(x_t|m'_{[0,t-1]}) P_\mu^\gamma(m'_{[0,t-1]}) \\
= &\sum_{(\mathcal{M}')^t} \sum_{\mathcal{X}} \sum_{\mathcal{M}'} ||\overline{\pi}_t^\mu - \overline{\pi}_t^\mu||_{TV} O_{Q_t}(m'_t|x_t)\pi_t^\mu(x) P_\mu^\gamma(m'_{[0,t-1]})
\end{aligned}$$

$$
\begin{aligned}
&\leq (2-\tilde{\delta}(O)) \sum_{(\mathcal{M}')^t} ||\pi_t^{\mu} - \pi_t^{\nu}||_{TV} P_{\mu}^{\gamma}(m'_{[0,t-1]}) \\
&= (2-\tilde{\delta}(O)) \mathbf{E}_{\mu}^{\gamma}\left[||\pi_t^{\mu} - \pi_t^{\nu}||_{TV}\right],
\end{aligned}
$$

where the third equality follows from the fact that Q_t is deterministic given μ and $M'_{[0,t-1]}$, and that given X_t and Q_t, M'_t depends only on the kernel O_{Q_t}. For the inequality we used Lemma 4.4.1. Finally, using the Dobrushin contraction property for kernels (as noted in the derivation of (4.19)), we have

$$
\begin{aligned}
\mathbf{E}_{\mu}^{\gamma}\left[||\pi_{t+1}^{\mu} - \pi_{t+1}^{\nu}||_{TV}\right] &\leq (1-\delta(T))\mathbf{E}_{\mu}^{\gamma}\left[||\overline{\pi}_t^{\mu} - \overline{\pi}_t^{\mu}||_{TV}\right] \\
&\leq (1-\delta(T))(2-\tilde{\delta}(O))\mathbf{E}_{\mu}^{\gamma}\left[||\pi_t^{\mu} - \pi_t^{\nu}||_{TV}\right].
\end{aligned}
$$

Chapter 5

Comparison: Belief Quantization vs Sliding Finite Window

In this section, we provide a detailed comparison noting explicit benefits and drawbacks of the schemes in the previous chapters. We also provide supporting simulations to illustrate some of our key points, as well as to generally show the performance of the algorithms in the previous chapters.

Unique invariant measure and convergence of Q-learning

- *Belief quantization (noiseless):* In Theorem 2.2.8, the convergence of V_t happens regardless of the initial distribution used during learning. This is due to the strong recurrence property noted in Lemma 2.2.1, which guarantees that the underlying $(\pi_t)_{t\geq 0}$ process will converge to its unique invariant distribution ϕ from any initialization. This property also made it straightforward to show the uniqueness of ϕ (see Lemma 2.2.3).

- *Belief quantization (noisy):* Since we lack a recurrence property, proving the uniqueness of ϕ was more difficult, and it was necessary to use asymptotic predictor stability. Accordingly, the convergence to ϕ (and thus, the convergence of V_t in Theorem 3.2.15) only holds for certain initial distributions (in particular, for ϕ-almost every initial distribution). However, since we do not know the exact form of ϕ, we relied on weak continuity and the additional Assumption 3.2.8 to claim convergence for $\pi_0 = \zeta$.
- *Sliding finite window:* Since the sliding finite window approximation is constant over $\mathcal{P}(\mathcal{X})$, the convergence to a unique invariant measure no longer depends on the (complex) properties of $(\pi_t)_{t\geq 0}$, and is instead inherited directly from the unique invariance of $(X_t)_{t\geq 0}$. Furthermore, since there are only finitely many finite windows, convergence in Theorem 4.2.4 holds for almost any initial window (under any prior).

Initialization during implementation of learned policy

- *Belief quantization (noiseless):* Again due to the recurrence property in Lemma 2.2.1, the near-optimality of the resulting policy holds for any of these recurrent elements (in particular, when π_0 is any row of T, as noted in the proof of Lemma 2.2.1).

- *Belief quantization (noisy):* Since the learned policy is only optimal over $\mathcal{P}_N^{\phi}(\mathcal{X})$, we require that the initial distribution during implementation is in the support of ϕ (in particular, we fix it to ζ). Note that we also require that the tie-breaking rule on the nearest neighbor map takes a certain form to ensure that we remain in $\mathcal{P}_N^{\phi}(\mathcal{X})$, as noted in Remark 3.2.17.

- *Sliding finite window:* Although we place restrictions on the source and channel in this scheme (to be discussed shortly), the initialization w_0 can be almost any finite window.

Conditions on the source and channel

- *Belief quantization:* Theorems 2.2.8 and 3.2.15 hold as long as the source admits a unique invariant distribution (though with the assumptions noted above on initializations).

- *Sliding finite window:* In order to ensure that the learned policy becomes near-optimal, we require a bound on the loss term L_t. In particular, we enforce that $\alpha := (1 - \delta(T))(2 - \delta(O)) < 1$, which means our sliding finite window approach is only valid on a restricted class of sources and channels.

Computational complexity and need for Bayesian updates

- *Belief quantization:* In both the noiseless and noisy setups, one must compute the true belief process $(\pi_t)_{t\geq 0}$ using the update equations, then quantize this to the set $\mathcal{P}_N(\mathcal{X})$. This increases the computational complexity of this scheme (and requires explicit knowledge of the system model), both during the Q-learning algorithm and during implementation of the learned policy.

- *Sliding finite window:* The sliding finite window scheme uses the approximate predictor from a fixed prior and a given history. Since there are only finitely many such histories, one can compute these offline before running the Q-learning and before implementation of the learned policy. They can then be accessed in a lookup table fashion, saving considerable computation especially when the alphabets are large.

Encoder/decoder implementation

- *Belief quantization:* In the belief quantization scheme, both the encoder and the decoder must track the true belief π_t. The encoder needs π_t to compute the proper value of $\hat{\pi}_t$ and apply the learned policy, while the decoder needs it to compute the optimal reproduction value $\hat{x}_t$.

- *Sliding finite window:* In the sliding finite window scheme the encoder policy is a constant function in $\mathcal{P}(\mathcal{X})$, so it can directly use the sliding finite window to apply the learned policy. The decoder must still compute $\hat{\pi}_t$ in order to calculate the reproduction value $\hat{x}_t$, however as in the previous point this can be done in a lookup table fashion. Note that in this case the decoder used is not strictly optimal for the encoder (as it uses the approximate $\hat{\pi}_t$ to compute $\hat{x}_t$), but under the Dobrushin coefficient conditions it is near-optimal for large N.

Rate of convergence to near-optimality

- *Belief quantization:* The near-optimality in Theorems 2.2.8 and 3.2.15 is only asymptotic as $N \to \infty$.

- *Sliding finite window:* For the sliding finite window result in Theorem 4.2.4, the convergence is exponential (note that although we only bound the expectation, since there are only finitely many initial memories, the convergence is also exponential in N for each initial finite window).

Quantization

- *Belief quantization:* To apply Theorem 1.4.5 we only require that $\max_\pi d(\pi, \hat{\pi}) \to 0$; accordingly, the quantization of $\mathcal{P}(\mathcal{X})$ does not have to be uniform. Theoretically, this allows a more efficient quantization, although in our implementation we always use the nearest neighbor scheme over $\mathcal{P}_N(\mathcal{X})$ (which gives a uniform quantization).

- *Sliding finite window:* The sliding finite window can be seen as a non-

	Belief quantization	Sliding finite window
Convergence of Q-learning and near-optimality	✓	✓
General source and channel	✓	✗
Insensitive to initialization	✗	✓
Exponential convergence of performance	✗	✓
Bayesian update not needed	✗	✓
Lookup table implementation	✗	✓

Table 5.1: Comparison of the two approximation schemes

uniform quantization of $\mathcal{W}$ (since it is constant over the belief coordinate). However, it is uniform over the product space $(\mathcal{M}')^N \times \mathcal{Q}^N$, since the sliding finite window scheme uses every possible finite window.

We summarize some of the key differences between the schemes in Table 5.1, considering both the noiseless and noisy channel belief quantization schemes together.

5.1 Simulations

We now give some examples of zero-delay coding problems and simulate the performance of the policies resulting from Theorems 2.2.8, 3.2.15, and 4.2.4. In all of the following, we use a discount factor $\beta = 0.9999$ and the distortion function $d(x, \hat{x}) = (x - \hat{x})^2$. The average distortion is calculated over $t = 0$ to $t = 10^5$. The initial distribution is ζ (that is, the invariant distribution for the source).

Remark 5.1.1. On implementation of the Q-learning algorithms: the theoretical upper bound for the possible number of states in each scheme grows very quickly in their respective parameters. In particular, for the belief quantization approach we have $|\mathcal{P}_N(\mathcal{X})| = \binom{N+|\mathcal{X}|+1}{|\mathcal{X}|-1}$ [45], and for the sliding finite window approach we have $|\mathcal{W}_N| = |\mathcal{M}' \times \mathcal{Q}|^N$. However, the sets actually visited during Q-learning (that is, $\mathcal{P}_N^{\phi}$ and $\mathcal{W}_N^{+}$) may be much smaller. Thus, one may wish to add entries to V_t as they are visited by the Q-learning algorithm in a dynamic fashion. Furthermore, note that for certain problems it may be possible to significantly shrink the set of quantizers $\mathcal{Q}$ with no loss of optimality. For example, for a noiseless channel one can omit those quantizers with empty bins, or for an i.i.d. source those with non-convex bins.

5.1.1 Comparison with Lloyd-Max quantizer for i.i.d. source and noiseless channel

Let $\mathcal{X} = \{0, \ldots, 7\}$ and consider an i.i.d. source $(X_t)_{t \geq 0}$, such that for all x,

$$T(\cdot|x) = \begin{pmatrix} \frac{1}{4} & \frac{1}{8} & \frac{1}{8} & \frac{1}{16} & \frac{1}{16} & \frac{1}{16} & \frac{1}{4} & \frac{1}{16} \end{pmatrix}.$$

Note that in the i.i.d. case, we trivially have $\delta(T) = 1$, so the sliding finite window approach is applicable. Indeed, here we have that $\alpha = 0$, so that for all $N \geq 1$,

$$\mathbf{E}^{\gamma}_{\pi_{-N}} \left[\left| J_\beta(w_0, \widetilde{\gamma}^*) - J^*_\beta(w_0) \right| \right] = 0.$$

That is, the optimal policy for the sliding finite window representation is optimal (not just near-optimal) for the original problem for any N. This is not surprising given the i.i.d. nature of the source; the approximation of π_{t-N} to ζ is without any loss since π_{t-N} can be immediately recovered.

Similarly for the quantization approach, $N = 1$ is sufficient since $\pi_t = \zeta$ for all $t \geq 0$, so increasing N does not change the resulting policy. Accordingly, we let $N = 1$ and compare the performance of both approaches against a Lloyd-Max quantizer when the channel is noiseless. We plot the performance for $N = 1$ and several sizes of $\mathcal{M}$. The rate is calculated as $\log_2 |\mathcal{M}|$. As expected by the above discussion, our algorithm matches with this quantizer in each case, which can be seen in Figure 5.1.

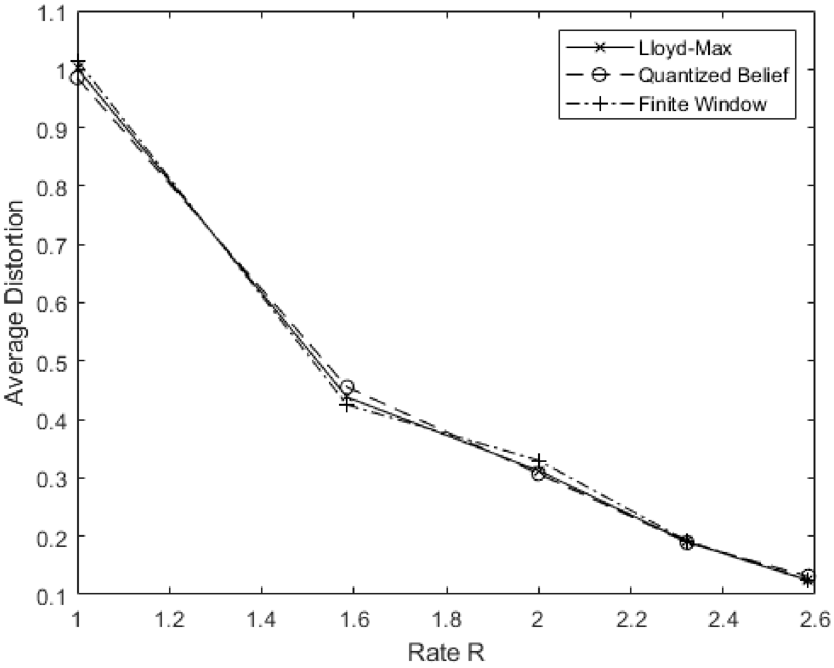

Figure 5.1: Comparison with Lloyd-Max

5.1.2 Comparison with memoryless encoding

Consider now a Markov source with transition kernel given by

$$T = \begin{pmatrix} \frac{1}{2} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}$$

and where the channel is a 4-ary symmetric channel with error probability 0.06:

$$O = \begin{pmatrix} 0.94 & 0.02 & 0.02 & 0.02 \\ 0.02 & 0.94 & 0.02 & 0.02 \\ 0.02 & 0.02 & 0.94 & 0.02 \\ 0.02 & 0.02 & 0.02 & 0.94 \end{pmatrix}.$$

We have that $\delta(T) = \frac{2}{3} > \frac{1}{2}$, so we can apply the finite memory scheme. In such a setup (where $\mathcal{X} = \mathcal{M}$ and the channel is symmetric), it was shown in [16] that "memoryless" encoding (that is, where $M_t = X_t$) is optimal. We compare our algorithms against such an encoding policy, shown in Figures 5.2 and 5.3, and note that both approach the optimum as N increases. Recall that N represents *different* parameters in the different approximation schemes; for the quantized belief method, it represents the common denominator of the finite set $\mathcal{P}_N(\mathcal{X})$, while it represents the window length for the sliding finite window method. Accordingly, we present the schemes on different sets of axes. Note that the convergence of the sliding finite window scheme to the optimum indeed appears exponential in N, as expected by Theorem 4.2.4.

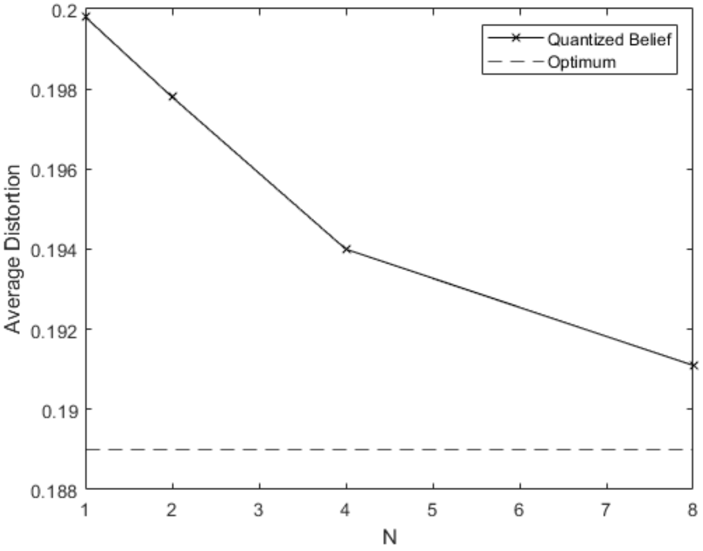

Figure 5.2: Quantized belief scheme vs memoryless encoding

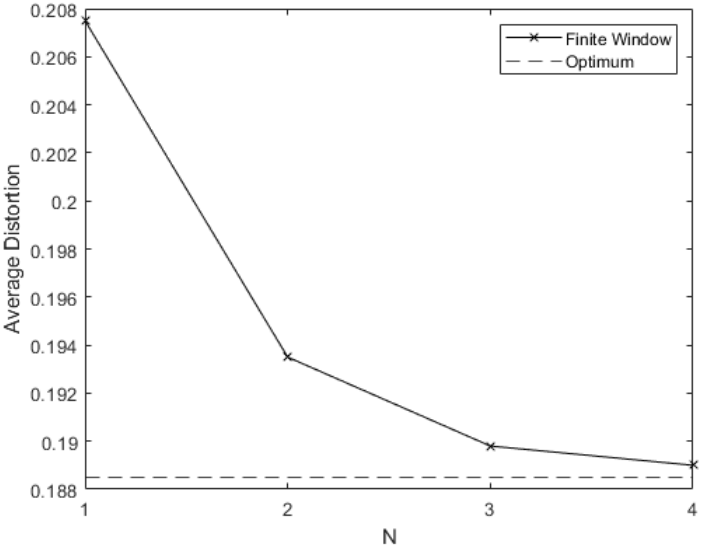

Figure 5.3: Finite memory scheme vs memoryless encoding

5.1.3 Problem with unknown optimum

Finally, we consider a setup where an optimal encoding scheme is unknown. Here we have a Markov source with transition kernel given by

$$T = \begin{pmatrix} 0.2476 & 0.1527 & 0.0775 & 0.2219 & 0.2082 & 0.0920 \\ 0.0805 & 0.0247 & 0.0776 & 0.1290 & 0.3718 & 0.3164 \\ 0.1510 & 0.2335 & 0.2042 & 0.0107 & 0.1425 & 0.2580 \\ 0.0056 & 0.2252 & 0.2303 & 0.2173 & 0.1141 & 0.2076 \\ 0.1357 & 0.2685 & 0.0494 & 0.1981 & 0.2930 & 0.0553 \\ 0.2373 & 0.2795 & 0.0698 & 0.0399 & 0.1371 & 0.2363 \end{pmatrix}.$$

Note that this transition kernel was randomly generated from the set of 6×6 transition matrices. The channel is a 3-ary symmetric channel with error probability 0.04:

$$O = \begin{pmatrix} 0.96 & 0.02 & 0.02 \\ 0.02 & 0.96 & 0.02 \\ 0.02 & 0.02 & 0.96 \end{pmatrix}.$$

It can be verified that the Dobrushin coefficient conditions of Theorem 4.2.4 are met for this setup. Figures 5.4 and 5.5 give the performance of both schemes for this problem.

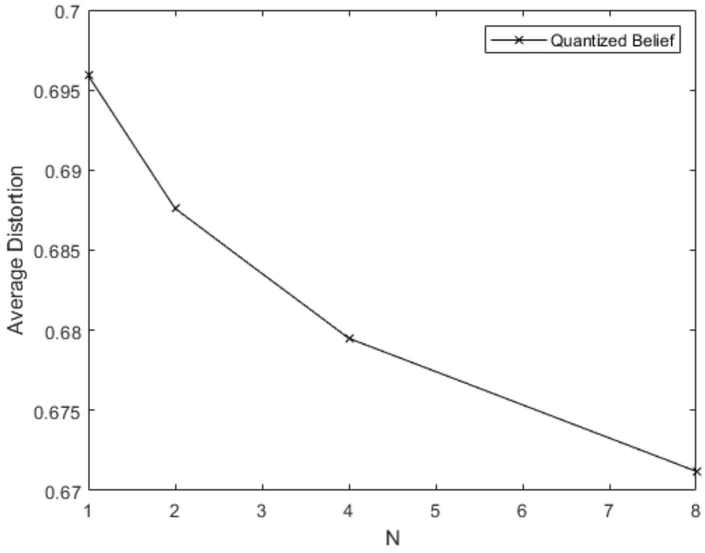

Figure 5.4: Quantized belief scheme with unknown optimum

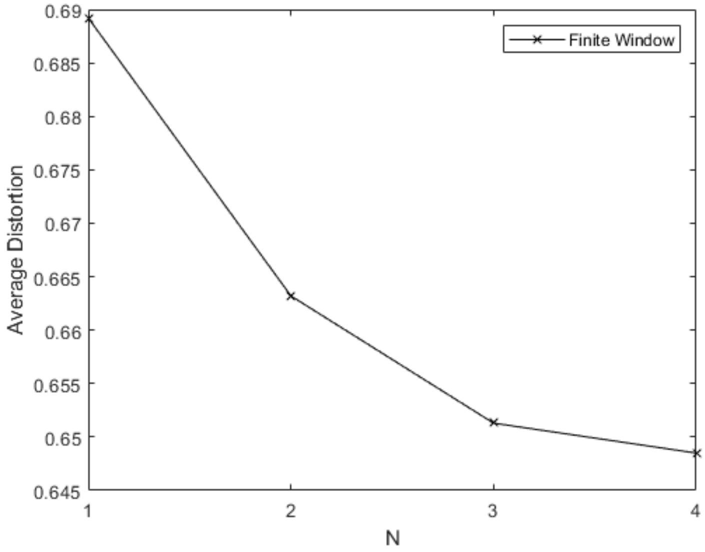

Figure 5.5: Finite memory scheme with unknown optimum

Chapter 6

Conclusion

6.1 Summary

In this thesis, we establish approximation and reinforcement learning results for the zero-delay coding problem over a possibly noisy channel with feedback, building on recent results in the stochastic control literature. The results crucially rely on the fact that the zero-delay coding problem can be formulated as a Markov decision process (MDP) with a belief-valued state and quantizer-valued control. This yields, to our knowledge, the first concrete implementation of an algorithm to find a provably near-optimal policy for this problem.

In Chapter 2, we review the results of [1] which examined the special case of a noiseless channel and approximate the problem by quantizing the underlying probability space. We leverage the fact that, under a policy which chooses quantizers uniformly, the true belief process admits a recurrent state from any initial prior. This is used to prove certain ergodic behavior of the approximate process, leading to the convergence of a Q-learning algorithm to compute a near-optimal policy.

In Chapter 3, we study the same approximation scheme, but this time with a possibly noisy channel. In this case, we lose the recurrent states from the noiseless setup, and instead rely on asymptotic filter stability to prove the ergodic behavior of the approximate process. As a consequence, there are some additional assumptions placed on the problem, especially on the initial distribution.

In Chapter 4, we instead use a sliding finite window of past observations to approximate the belief term, necessitating an alternative MDP formulation and the development of several supporting results, which generalize existing results on filter stability to our setup. We show that this scheme results in exponential convergence to near-optimality and insensitivity to initialization, but only for a restricted class of sources and channels.

Finally, in Chapter 5 we discuss several important differences between the schemes and provided supporting simulation results, comparing our policies against optimal policies, when these are known. When they are not known, we show convergent behaviour as expected by the rigorous near-optimality results earlier in the thesis.

6.2 Future Work

There are several promising research directions for this problem. Firstly, a generalization to continuous alphabets; the approximations would then be done by approximating the continuous source alphabet by a compact one, then by a finite one, and finally applying similar approximation techniques to those used in this thesis. In [53] and [54, Section V.C], under certain assumptions, this was shown to be an efficient method for quantizing probability measures under a Wasserstein metric, and consequently, the weak convergence topology. The ergodic and weak continuity properties are expected to follow by similar methods to those used here. However, an important distinction is that the cost function is now unbounded - this requires additional analysis as the key MDP theorems used here assume a bounded cost function.

We also wish to find less stringent filter stability conditions for the sliding finite window scheme than the Dobrushin coefficient ones given here; this may be possible via observability-type conditions, such as those in [55]. However, the dependence on past quantizers makes this analysis difficult.

Finally, a very practical generalization is the controlled case, in which one wishes to send information over a channel to a controller who takes some control action. Many of the results here are expected to carry over to this case, with a modification of the MDP formulation.

Bibliography

[1] L. Cregg, T. Linder, and S. Yüksel, "Reinforcement learning for near-optimal design of zero-delay codes for Markov sources," *IEEE Transactions on Information Theory (also arXiv:2311.12609)*, 2024.

[2] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[3] E. I. Silva, M. S. Derpich, and J. Østergaard, "A framework for control system design subject to average data-rate constraints," *IEEE Transactions on Automatic Control*, vol. 56, pp. 1886–1899, 2011.

[4] E. I. Silva, M. S. Derpich, J. Østergaard, and M. Encina, "A characterization of the minimal average data rate that guarantees a given closed-loop performance level," *IEEE Transactions on Automatic Control*, vol. 61, no. 8, pp. 2171–2186, 2015.

[5] P. A. Stavrou, J. Østergaard, and C. D. Charalambous, "Zero-delay rate distortion via filtering for vector-valued Gaussian sources," *IEEE*

Journal of Selected Topics in Signal Processing, vol. 12, no. 5, pp. 841–856, 2018.

[6] R. Bansal and T. Başar, “Simultaneous design of measurement and control strategies in stochastic systems with feedback,” *Automatica*, vol. 45, pp. 679–694, 1989.

[7] S. Tatikonda, A. Sahai, and S. Mitter, “Stochastic linear control over a communication channels,” *IEEE Transactions on Automatic Control*, vol. 49, pp. 1549–1561, 2004.

[8] T. Tanaka, K.-K. K. Kim, P. A. Parrilo, and S. K. Mitter, “Semidefinite programming approach to Gaussian sequential rate-distortion trade-offs,” *IEEE Transactions on Automatic Control*, vol. 62, no. 4, pp. 1896–1910, 2016.

[9] M. S. Derpich and J. Østergaard, “Improved upper bounds to the causal quadratic rate-distortion function for Gaussian stationary sources,” *IEEE Transactions on Information Theory*, vol. 58, no. 5, pp. 3131–3152, 2012.

[10] P. A. Stavrou and M. Skoglund, “Asymptotic reverse waterfilling algorithm of NRDF for certain classes of vector Gauss-Markov processes,” *IEEE Transactions on Automatic Control*, vol. 67, no. 6, pp. 3196–3203, 2022.

[11] P. A. Stavrou, T. Tanaka, and S. Tatikonda, “The time-invariant multidimensional Gaussian sequential rate-distortion problem revisited,”

IEEE Transactions on Automatic Control, vol. 65, no. 5, pp. 2245–2249, 2019.

[12] V. Kostina and B. Hassibi, "Rate-cost tradeoffs in control," *IEEE Transactions on Automatic Control*, vol. 64, no. 11, pp. 4525–4540, 2019.

[13] D. Pollard, "Quantization and the method of k-means," *IEEE Transactions on Information Theory*, vol. 28, pp. 199–205, 1982.

[14] T. Linder, G. Lugosi, and K. Zeger, "Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding," *IEEE Transactions on Information Theory*, vol. 40, no. 6, pp. 1728–1740, 1994.

[15] T. Linder, *Learning-theoretic methods in vector quantization.* Springer, Wien, New York, 2002, pp. 163–210.

[16] J. C. Walrand and P. Varaiya, "Optimal causal coding-decoding problems," *IEEE Transactions on Information Theory*, vol. 19, pp. 814–820, 1983.

[17] H. S. Witsenhausen, "On the structure of real-time source coders," *Bell Syst. Tech. J*, vol. 58, pp. 1437–1451, 1979.

[18] R. G. Wood, T. Linder, and S. Yüksel, "Optimal zero delay coding of Markov sources: Stationary and finite memory codes," *IEEE Transactions on Information Theory*, vol. 63, pp. 5968–5980, 2017.

[19] D. L. Neuhoff and R. K. Gilbert, “Causal source codes,” *IEEE Transactions on Information Theory*, vol. 28, pp. 701–713, 1982.

[20] N. Gaarder and D. Slepian, “On optimal finite-state digital transmission systems,” *IEEE Transactions on Information Theory*, vol. 28, pp. 167–186, 1982.

[21] T. Weissman and N. Merhav, “On causal source codes with side information,” *IEEE Transactions on Information Theory*, vol. 51, pp. 4003–4013, 2005.

[22] H. Asnani and T. Weissman, “On real time coding with limited lookahead,” *IEEE Transactions on Information Theory*, vol. 59, no. 6, pp. 3582–3606, 2013.

[23] E. Bourtsoulatze, D. Kurka, and D. Gündüz, “Deep joint source-channel coding for wireless image transmission,” *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 3, pp. 567–579, 2019.

[24] D. Gündüz, P. de Kerret, N. Sidiropoulos, D. Gesbert, C. Murthy, and M. van der Schaar, “Machine learning in the air,” *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 10, pp. 2184–2199, 2019.

[25] N. Farsad, M. Rao, and A. Goldsmith, “Deep learning for joint source-channel coding of text,” in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 2326–2330.

[26] E. Bourtsoulatze, D. Burth Kurka, and D. Gündüz, “Deep joint source-channel coding for wireless image transmission,” *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 3, pp. 567–579, 2019.

[27] Z. Aharoni, O. Sabag, and H. H. Permuter, “Computing the feedback capacity of finite state channels using reinforcement learning,” in *2019 IEEE International Symposium on Information Theory (ISIT)*, 2019, pp. 837–841.

[28] Z. Aharoni, O. Sabag, and H. H. Permuter, “Feedback capacity of Ising channels with large alphabet via reinforcement learning,” *IEEE Transactions on Information Theory*, vol. 68, no. 9, pp. 5637–5656, 2022.

[29] D. Teneketzis, “On the structure of optimal real-time encoders and decoders in noisy communication,” *IEEE Transactions on Information Theory*, vol. 52, pp. 4017–4035, 2006.

[30] A. Mahajan and D. Teneketzis, “Optimal design of sequential real-time communication systems,” *IEEE Transactions on Information Theory*, vol. 55, pp. 5317–5338, 2009.

[31] M. Ghomi, T. Linder, and S. Yüksel, “Zero-delay lossy coding of linear vector Markov sources: Optimality of stationary codes and near optimality of finite memory codes,” *IEEE Transactions on Information Theory*, vol. 68, no. 5, pp. 3474–3488, 2021.

[32] O. Hernandez-Lerma and J. B. Lasserre, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. New York: Springer, 1996.

[33] S. Yüksel, "Optimization and Control of Stochastic Systems," `https://mast.queensu.ca/~yuksel/LectureNotesOnStochasticOptControl.pdf`, 2023.

[34] G. Ungerboeck, "Channel coding with multilevel/phase signals," *IEEE Transactions on Information Theory*, vol. 28, no. 1, pp. 55–67, 1982. DOI: `10.1109/TIT.1982.1056454`.

[35] E. Ayanoglu and R. Gray, "The design of joint source and channel trellis waveform coders," *IEEE Transactions on Information Theory*, vol. 33, no. 6, pp. 855–865, 1987. DOI: `10.1109/TIT.1987.1057376`.

[36] J. Dunham and R. Gray, "Joint source and noisy channel trellis encoding (corresp.)," *IEEE Transactions on Information Theory*, vol. 27, no. 4, pp. 516–519, 1981. DOI: `10.1109/TIT.1981.1056366`.

[37] A. Viterbi and J. Omura, "Trellis encoding of memoryless discrete-time sources with a fidelity criterion," *IEEE Transactions on Information Theory*, vol. 20, no. 3, pp. 325–332, 1974. DOI: `10.1109/TIT.1974.1055233`.

[38] N. Saldi, T. Linder, and S. Yüksel, *Finite Approximations in Discrete-Time Stochastic Control: Quantized Models and Asymptotic Optimality*. Birkhäuser, 2018.

[39] A. D. Kara and S. Yüksel, “Q-learning for stochastic control under general information structures and non-Markovian environments,” *Transactions on Machine Learning Research, arXiv:2311.00123*, 2024.

[40] T. Linder and S. Yüksel, “On optimal zero-delay coding of vector Markov sources,” *IEEE Transactions on Information Theory*, vol. 60, pp. 2975–5991, 2014.

[41] O. Cappé, E. Moulines, and T. Rydén, *Inference in Hidden Markov Models*, English. Germany: Springer, 2005.

[42] P. Chigansky and R. Liptser, “Stability of nonlinear filters in non-mixing case,” *Annals of Applied Probability*, vol. 14, pp. 2038–2056, 2004.

[43] A. D. Kara, N. Saldi, and S. Yüksel, “Q-learning for MDPs with general spaces: Convergence and near optimality via quantization under weak continuity,” *Journal of Machine Learning Research*, vol. 24, no. 199, pp. 1–34, 2023.

[44] A. D. Kara and S. Yüksel, “Convergence of finite memory Q-learning for POMDPs and near optimality of learned policies under filter stability,” *Mathematics of Operations Research*, vol. 48, no. 4, pp. 2066–2093, 2023.

[45] Y. A. Reznik, “An algorithm for quantization of dicrete probability distributions,” *DCC 2011*, pp. 333–342, 2011.

[46] O. Hernández-Lerma and J. B. Lasserre, *Markov chains and invariant probabilities*. Basel: Birkhäuser-Verlag, 2003.

[47] O. Hernández-Lerma, R. Montes-de-Oca, and R. Cavazos-Cadena, "Recurrence conditions for markov decision processes with borel state space: A survey," *Annals of Operations Research*, vol. 28, no. 1, pp. 29–46, 1991.

[48] R. L. Dobrushin, "Central limit theorem for nonstationary Markov chains. i," *Theory of Probability & Its Applications*, vol. 1, no. 1, pp. 65–80, 1956.

[49] R. van Handel, "The stability of conditional Markov processes and Markov chains in random environments," *Annals of Applied Probability*, vol. 37, pp. 1876–1925, 2009.

[50] G. B. D. Masi and L. Stettner, "Ergodicity of hidden Markov models," *Mathematics of Control, Signals and Systems*, vol. 17, no. 4, pp. 269–296, 2005.

[51] M. Hairer. "Convergence of Markov processes." (2010), [Online]. Available: `https://www.hairer.org/notes/Convergence.pdf`.

[52] C. McDonald and S. Yüksel, "Exponential filter stability via Dobrushin's coefficient," *Electronic Communications in Probability*, vol. 25, 2020.

[53] W. Kreitmeier, "Optimal vector quantization in terms of Wasserstein distance," *Journal of Multivariate Analysis*, vol. 102, no. 8, pp. 1225–1239, 2011.

[54] N. Saldi, S. Yüksel, and T. Linder, "Finite model approximations for partially observed Markov decision processes with discounted cost," *IEEE Transactions on Automatic Control*, vol. 65, 2020.

[55] C. McDonald and S. Yüksel, "Stochastic observability and filter stability under several criteria," *IEEE Transactions on Automatic Control*, pp. 1–16, 2023.