Sliding Window Codes: Near-Optimality and Q-Learning for Zero-Delay Coding

Liam Cregg, Fady Alajaji, Serdar Yüksel

arXiv:2310.06742v4 [math.OC] 6 Oct 2025

Abstract

We study the problem of zero-delay coding for the transmission of a Markov source over a noisy channel with feedback and present a reinforcement learning solution which is 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, but their computation is prohibitively complex due to the uncountable nature of the constructed state space and the lack of minorization or strong ergodicity results. These challenges invite rigorous reinforcement learning methods, which entail several open questions: can we approximate this MDP with a finite-state one with some performance guarantee? Can we ensure convergence of a reinforcement learning algorithm for this approximate MDP? What regularity assumptions are required for the above to hold? We address these questions as follows: we present an approximation of the belief MDP using a sliding finite window of channel outputs and quantizers. Under an appropriate notion of predictor stability, we show that policies based on this finite window are near-optimal, in the sense that the lowest distortion achievable by such a policy approaches the true lowest distortion as the window length increases. We give sufficient conditions for predictor stability to hold. Finally, we propose a Q-learning algorithm which provably converges to a near-optimal policy and provide a detailed comparison of the sliding finite window scheme with another approximation scheme which quantizes the belief MDP in a nearest neighbor fashion.

Index Terms

Zero-delay coding, reinforcement learning, Q-learning, stochastic control, Markov decision processes, quantization, source-channel coding, noisy channels with feedback

I. Introduction

The zero-delay coding problem involves compressing and transmitting an information source at a fixed rate over a noisy channel with feedback and without delay, while minimizing the expected distortion at the receiver. This zero-delay restriction is of practical relevance in many applications, including live-streaming [1], [2], [3], [4], real-time tracking and estimating processes over erasure links [5] and real-time sensor networks [6], [7]. However, this

This work was supported by the Natural Sciences and Engineering Research Council of Canada. The material in this paper has been presented in part (without the associated technical analysis) at the 2023 Conference on Decision and Control and the 2024 American Control Conference.

The first author is with the Department of Information Technology and Electrical Engineering, ETH Zürich, Switzerland. The second and third authors are with the Department of Mathematics and Statistics, Queen's University, Kingston, Canada (e-mail: lcregg@ethz.ch; fa@queensu.ca; yuksel@queensu.ca).

restriction means that classical Shannon-theoretic methods [8], which require collecting large sequences of source symbols and compressing them at once, are not viable as they induce a large delay. 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 [9], [10], [11]. Within the context of linear systems, [12], [13], [14], [15], [16] 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 [17] and [18] for further studies, and see [19], [20], [21] for studies in the context of channels with delay and other communication constraints.

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 [22], [23], [24] 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 [25], [26], [27] (see also, for related discussions in a different causal coding context [28], [29], [30], [31]). More recently, deep learning is employed to construct powerful joint source-channel codes (see [32], [33], [34]), and reinforcement learning is used as a tool to estimate feedback capacity in [35], [36]. 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*, in the sense that, for any $\epsilon > 0$, there exists some window size N such that the resulting policy achieves distortion no further than ϵ from the optimal distortion (a formal definition is given in Definition 2.3).

There have been several studies about the zero-delay coding problem using stochastic control techniques. In particular, [26], [25], [37] 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 [27] (with feedback) and [38] (without feedback). The continuous-alphabet infinite-horizon case is examined in [39], 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 [40], [41] 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.

In this paper, we present an approximation method to simplify the resulting MDP, and we use this approximation 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 [42], which were originally used to study partially observed Markov decision processes (POMDPs). Based on these methods, we introduce a practical sliding finite window method and present several mathematical and

algorithmic results on its near-optimal performance. We also compare this method with the approach used in [43], which uses a nearest neighbor quantization of the probability measure-valued state space. Although [43] only studied the noiseless channel (quantization) version of the problem, extensions to the noisy channel case follow with little additional effort as we discuss in the paper.

We emphasize that our new sliding finite window method has several advantages over the belief quantization scheme. In particular, it is easier to implement, computationally less complex, and valid for any initialization. This comes at a cost of some additional Dobrushin coefficient conditions on the source and channel, but we note that the conditions provided here are sufficient, not necessary. These differences can be seen by comparing Theorems 4.1 and 5.2, and are further detailed in Section V.

The sliding finite window code that we propose can be viewed as a finite state code [44], [45] where the states are the realizations of a finite window of encoder maps and outputs. We are not aware of such a construction in the literature with rigorous performance guarantees in the context of optimal real-time coding and in particular on an associated learning theoretic study. We also note the similarities to trellis or convolutional codes [46], [47], [48]; our method can be in some sense be seen as a zero-delay analog of these codes, but our analysis is based on stochastic control and thus entirely different than that of trellis codes.

Main Contributions

- To our knowledge there is no prior study on near-optimality of sliding window coding schemes for the zero-delay coding problem with an explicit rate of convergence (involving all window lengths). To this end, in Theorem 3.4, we provide an approximation result for the zero-delay coding problem which bounds the sub-optimality gap of sliding finite window policies by a predictor stability term.
- Under Dobrushin coefficient conditions, we then show that these sliding finite window policies become near-optimal for the zero-delay coding problem in Corollary 3.7, with an explicit performance bound. We note that these conditions hold when the channel is "noisy enough" or when the source has some mixing conditions. To our knowledge, this gives the first rigorous proof of optimality of sliding finite window policies for the zero-delay coding problem.
- We provide a (reinforcement) Q-learning algorithm which allows for the computation of these sliding finite window policies and rigorously show its convergence in Theorem 4.1.
- Finally, we provide both theoretical and experimental comparisons of our sliding finite window method with a nearest neighbor approximation scheme presented in [43] (and adapted and generalized to the noisy channel setting).

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 $\mathcal{Y}$.

II. Preliminaries: Optimal Coding Problem and its MDP Formulation

A. Optimal Zero-Delay Coding and Existence of an Optimal Policy

We recall a few basic definitions and results among the large body of research on Markov chains [49]. Let $(a_t)_{t\geq 0}$ be an A-valued Markov chain, where $(A, \mathcal{B})$ is a measurable space. Denote its transition kernel by $K(B|a) \coloneqq P(a_{t+1} \in B|a_t = a)$ for any measurable set $B \in \mathcal{B}$, and its t-fold transition kernel by

$$K^t(B|a) \coloneqq \int_{A^{t-1}} K(da_1|a)K(da_2|a_1)\dots K(da_{t-1}|a_{t-2})K(B|a_{t-1}). \tag{II.1}$$

Definition 2.1: We say a probability measure μ is *invariant* for kernel K (equivalently, for the Markov chain $(a_t)_{t\geq 0}$) if

$$\int_A K(B|a)\mu(da) = K(B) \text{ for all } B \in \mathcal{B}. \tag{II.2}$$

Definition 2.2: Let ψ be a nontrivial σ-finite measure over $\mathcal{B}$. Then $(a_t)_{t\geq 0}$ is ψ-*irreducible* if

$$\sum_{t\geq 0} K^t(B|a) > 0 \text{ for all } a \in A \text{ whenever } \psi(B) > 0. \tag{II.3}$$

When A is countable, we let ψ be the counting measure and simply call the Markov chain irreducible. One classical result in Markov chain theory is that an irreducible Markov chain on a finite space has a unique invariant measure [49].

Let our information source be a Markov process $(X_t)_{t\geq 0}$ taking values in $\mathcal{X}$, which we assume is finite. Let $T(x'|x) \coloneqq P(X_{t+1} = x'|X_t = x)$ be its transition kernel, which we assume is irreducible, and thus admits a unique invariant measure. Denote this unique invariant measure by ζ. Let $X_0 \sim \pi_0$ (we also call π_0 the prior). Let $\mathcal{M}$ and $\mathcal{M}'$ be the input and output alphabets of the memoryless noisy channel, which we assume are finite, and let $(M_t)_{t\geq 0}$ and $(M'_t)_{t\geq 0}$ be the respective processes. We denote the channel kernel by $O(m'|m) \coloneqq P(M'_t = m'|M_t = m)$ for $m' \in \mathcal{M}'$ and $m \in \mathcal{M}$. Finally, let $\hat{\mathcal{X}}$ be some finite set of reconstruction values, and let $(\hat{X}_t)_{t\geq 0} \subset \hat{\mathcal{X}}$ be the corresponding sequence of reproductions.

Consider sequences of functions $(\gamma^e_t)_{t\geq 0}$, which we call the encoder policy, and $(\gamma^d_t)_{t\geq 0}$, 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

$$\gamma^e_t : \mathcal{X}^{t+1} \times \mathcal{M}^t \times (\mathcal{M}')^t \to \mathcal{M} \qquad \gamma^d_t : (\mathcal{M}')^{t+1} \to \hat{\mathcal{X}} \tag{II.4}$$

$$(X_{[0,t]}, M_{[0,t-1]}, M'_{[0,t-1]}) \mapsto M_t \qquad M'_{[0,t]} \mapsto \hat{X}_t. \tag{II.5}$$

Our setup can be seen in Figure 1. Note that the initial distribution π_0 and the policies $(\gamma^e_t, \gamma^d_t)_{t\geq 0}$ induce a joint distribution on $(X_t, M_t, M'_t, \hat{X}_t)_{t\geq 0}$; this follows from the Ionescu-Tulcea Theorem, see e.g., [40, Proposition C.10].

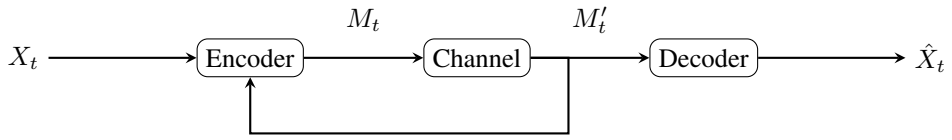


Fig. 1: Source-channel coding with feedback

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,

$$J_\beta(\pi_0, \gamma^e, \gamma^d) \coloneqq \mathbf{E}_{\pi_0}^{\gamma^e, \gamma^d}\left[\sum_{t=0}^{\infty} \beta^t d(X_t, \hat{X}_t)\right], \tag{II.6}$$

or the average distortion,

$$J(\pi_0, \gamma^e, \gamma^d) \coloneqq \limsup_{T\to\infty} \mathbf{E}_{\pi_0}^{\gamma^e, \gamma^d}\left[\frac{1}{T}\sum_{t=0}^{T-1} d(X_t, \hat{X}_t)\right], \tag{II.7}$$

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 refer to the minimization of (II.6) as the discounted distortion problem and of (II.7) 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

$$\hat{X}_t = \gamma_t^{d*}(M'_{[0,t]}) \coloneqq \operatorname*{argmin}_{\hat{x}\in\hat{\mathcal{X}}} \mathbf{E}_{\pi_0}^{\gamma^e}\left[d(X_t, \hat{x})|M'_{[0,t]}\right]. \tag{II.8}$$

Accordingly, we assume that we use an optimal decoder policy for a given encoder policy. We then denote by Γ the set of all encoder policies, and refer to the resulting encoder-decoder policy (γ^e, γ^{d*}) simply as γ. Then (II.6) and (II.7) become $J_\beta(\pi_0, \gamma)$ and $J(\pi_0, \gamma)$, respectively. We denote the minimal distortions by

$$J_\beta^*(\pi_0) \coloneqq \inf_\gamma J_\beta(\pi_0, \gamma) \tag{II.9}$$

$$J^*(\pi_0) \coloneqq \inf_\gamma J(\pi_0, \gamma). \tag{II.10}$$

We will also consider policies which obtain the above infima within some arbitrary threshold $\epsilon > 0$, which we call *near-optimal*, as follows:

Definition 2.3: 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, it can be shown 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 (see [41, Theorem 7.3.6]). Furthermore, the discounted distortion problem is generally easier to study from a reinforcement learning standpoint. Thus, we will target the discounted distortion problem throughout the majority of the paper and then make connections with the average distortion problem by taking $\beta \to 1$.

For fixed $(x_{[0,t-1]}, m_{[0,t-1]}, m'_{[0,t-1]})$, 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 [50].

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 $g_Q(x, m') \coloneqq \frac{dO_Q}{d\psi}(x, m')$ 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

$$\pi_t(\cdot) = P^{\gamma}_{\pi_0}(X_t \in \cdot | M'_{[0,t-1]}) \tag{II.11}$$

$$\overline{\pi}_t(\cdot) = P^{\gamma}_{\pi_0}(X_t \in \cdot | M'_{[0,t]}), \tag{II.12}$$

recalling that $X_0 \sim \pi_0$. We have dropped the 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*, as it is the predictive probability of the next symbol X_t. $\overline{\pi}_t$ is called the *filter* due to its role in estimation and nonlinear filtering of partially observable systems [51]. With a slight abuse of notation, we also let the source transition kernel T act as an operator on probability measures as follows:

$$T : \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{X}) \tag{II.13}$$

$$\pi(x) \mapsto \sum_{\mathcal{X}} T(x'|x)\pi(x). \tag{II.14}$$

Then given π_0, the above measures can be computed in a recursive manner as follows (see [52, Proposition 3.2.5]).

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

$$\pi_{t+1} = T(\overline{\pi}_t). \tag{II.15}$$

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 2.4: [27] 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 [25] for a finite time horizon and extended to the infinite-horizon case in [27].

Proposition 2.5: [27, Proposition 2] For any $\beta \in (0, 1)$, there exists $\gamma^* \in \Gamma_{\text{WS}}$ that solves the discounted distortion problem (that is, it minimizes (II.6)) for all priors $\pi_0 \in \mathcal{P}(\mathcal{X})$.

Proposition 2.6: [27, Theorem 3] There exists $\gamma^* \in \Gamma_{\text{WS}}$ that solves the average distortion problem (that is, it minimizes (II.7)) for all priors $\pi_0 \in \mathcal{P}(\mathcal{X})$.

B. Regularity Properties of the Markov Decision Process

Utilized in the above results is the fact that, under any $\gamma \in \Gamma_{\text{WS}}$, the zero-delay coding problem can be viewed as a Markov decision process (MDP), which we now formally define.

Definition 2.7: 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*.

The objective is to minimize $J_\beta(z_0, \gamma) \coloneqq \mathbf{E}^\gamma_{z_0}\left[\sum_{t=0}^\infty \beta^t c(Z_t, U_t)\right]$ (which we call the discounted cost problem) or $J(z_0, \gamma) \coloneqq \limsup_{T\to\infty} \mathbf{E}^\gamma_{z_0}\left[\frac{1}{T}\sum_{t=0}^{T-1} c(Z_t, U_t)\right]$ (which we call the average cost problem), over all γ, where $\gamma = (\gamma_t)_{t\geq 0}$ and $U_t = \gamma_t(Z_{[0,t]}, U_{[0,t-1]})$.

Proposition 2.8: Under any $\gamma \in \Gamma_{\text{WS}}$, the zero-delay coding problem is an MDP, where:

1) $\mathcal{Z} = \mathcal{P}(\mathcal{X})$.
2) $\mathcal{U} = \mathcal{Q}$.
3) $P = P(d\pi'|\pi, Q)$ induced by the update equations in (II.15).
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 (II.15) and the fact that, under any $\gamma \in \Gamma_{\text{WS}}$, π_t completely determines Q_t. The choice of c is due to the following result.

Lemma 2.9: If an optimal decoder is used, the expected distortion at the encoder (that is, before sending M_t) is given by

$$c(\pi_t, Q_t) = \sum_{\mathcal{M}'} \min_{\hat{x}\in\hat{\mathcal{X}}} \sum_{\mathcal{X}} d(x, \hat{x}) O_{Q_t}(m'|x)\pi_t(x). \tag{II.16}$$

Proof: Note that, under any $\gamma \in \Gamma_{\text{WS}}$ and given π_0, the optimal decoder can compute the filter $\overline{\pi}_t$, and thus the optimal decoder chooses $\hat{X}_t$ according to

$$\operatorname*{argmin}_{\hat{x}} \mathbf{E}\Big[d(X_t, \hat{x})|M'_{[0,t]}\Big] = \operatorname*{argmin}_{\hat{x}} \sum_x d(x, \hat{x})\overline{\pi}_t(x). \tag{II.17}$$

By the update equations in (II.15), we have

$$\overline{\pi}_t(x) = \frac{g_{Q_t}(x, M'_t)\pi_t(x)}{N_t(M'_t, Q_t)}, \tag{II.18}$$

where $N_t(m', Q) = \sum_{\mathcal{X}} g_Q(x, m')\pi_t(x)$. Thus at the decoder, the expected distortion is given by

$$\min_{\hat{x}} \sum_x d(x, \hat{x}) \frac{g_{Q_t}(x, M'_t)\pi_t(x)}{N_t(M'_t, Q_t)}. \tag{II.19}$$

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

$$\sum_{m'} \min_{\hat{x}} \sum_x d(x, \hat{x}) \frac{g_{Q_t}(x, m')\pi_t(x)}{N_t(m', Q_t)} P^\gamma_{\pi_0}(M'_t = m'|M'_{[0,t-1]}) \tag{II.20}$$

$$= \sum_{m'} \min_{\hat{x}} \sum_x d(x, \hat{x}) \frac{g_{Q_t}(x, m')\pi_t(x)}{N_t(m', Q_t)} \sum_x P^\gamma_{\pi_0}(M'_t = m', X_t = x|M'_{[0,t-1]}) \tag{II.21}$$

$$= \sum_{m'} \min_{\hat{x}} \sum_{x} d(x,\hat{x}) \frac{g_{Q_t}(x,m')\pi_t(x)}{N_t(m',Q_t)} \sum_{x} P^{\gamma}_{\pi_0}(X_t = x | M'_{[0,t-1]}) P^{\gamma}_{\pi_0}(M'_t = m' | X_t = x, M'_{[0,t-1]}) \tag{II.22}$$

$$= \sum_{m'} \min_{\hat{x}} \sum_{x} d(x,\hat{x}) \frac{g_{Q_t}(x,m')\pi_t(x)}{N_t(m',Q_t)} \sum_{x} \pi_t(x) O_{Q_t}(m'|x) \tag{II.23}$$

$$= \sum_{m'} \min_{\hat{x}} \sum_{x} d(x,\hat{x}) O_{Q_t}(m'|x)\pi_t(x), \tag{II.24}$$

where third equality holds by the definition of π_t and since $P^{\gamma}_{\pi_0}(M'_t = m'|X_t = x, M'_{[0,t-1]}) = O_{Q_t}(m'|x)$ by the memoryless property of the channel. The last equality follows since $\sum_x \pi_t(x) O_{Q_t}(m'|x) = |\mathcal{M}'|^{-1} N_t(m', Q_t)$ and $g_{Q_t}(x,m') = |\mathcal{M}'| O_{Q_t}(m'|x)$. ■

By this lemma, we have that the expected distortion at the encoder (assuming an optimal decoder), satisfies

$$\mathbf{E}^{\gamma}_{\pi_0}\left[\sum_{t=0}^{T-1} c(\pi_t, Q_t)\right] = \mathbf{E}^{\gamma}_{\pi_0}\left[\sum_{t=0}^{T-1} d(X_t, \hat{X}_t)\right]. \tag{II.25}$$

Thus, this choice of c ensures that solving the MDP defined in Proposition 2.8 over all $\gamma \in \Gamma_{\text{WS}}$ (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 for this MDP (rather than the original 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 finite and irreducible (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. We will cover this method in detail in Section IV. Finally, although explicit computation of $P(d\pi'|\pi, Q)$ is difficult, the following key structural result was obtained in [50].

Lemma 2.10: [50, Lemma 11] The transition kernel $P(d\pi'|\pi, Q)$ is weakly continuous. That is,

$$\int_{\mathcal{P}(\mathcal{X})} f(\pi') P(d\pi'|\pi, Q) \tag{II.26}$$

is continuous in (π, Q) for any continuous and bounded $f : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$.

Here, we endow $\mathcal{P}(\mathcal{X})$ with the weak convergence topology and $\mathcal{Q}$ with the Young topology (see [50]). Alternatively, since $\mathcal{Q}$ is finite here the discrete topology would also suffice. We note that MDPs with weakly continuous transition kernels as above are often called *weak Feller*.

C. Filter and Predictor Stability

A key property that we use is *filter/predictor* stability (recall from Definition II.11 that the predictor is given by π_t and the filter by $\overline{\pi}_t$).

Definition 2.11: The total variation distance between two probability measures μ, ν defined over $\mathcal{X}$ is given by

$$||\mu - \nu||_{\text{TV}} \coloneqq \sup_{||f||_\infty \leq 1} \left| \int_{\mathcal{X}} f(x)\mu(dx) - \int_{\mathcal{X}} f(x)\nu(dx) \right|, \tag{II.27}$$

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 (II.15) and note that they are sensitive to the value of π_0. Accordingly, we use π_t^μ to denote the predictor when $\pi_0 = \mu$.

Definition 2.12: The predictor process $(\pi_t)_{t\geq 0}$ is *stable in total variation almost surely* under a policy $\gamma \in \Gamma_{\text{WS}}$ if, for all $\mu, \nu \in \mathcal{P}(\mathcal{X})$ such that $\mu \ll \nu$, we have

$$\lim_{t\to\infty} ||\pi_t^\mu - \pi_t^\nu||_{\text{TV}} = 0 \; P_\mu^\gamma\text{-a.s.}, \tag{II.28}$$

where π_t^μ (respectively, π_t^ν) denotes the predictor with prior $\pi_0 = \mu$ (respectively, $\pi_0 = \nu$).
That is, the predictor process is insensitive to its initialization. In some cases, we are interested in a different form of stability, which we define next.

Definition 2.13: The predictor process $(\pi_t)_{t\geq 0}$ is *exponentially stable in total variation in expectation* under a policy $\gamma \in \Gamma_{\text{WS}}$ if there exists a coefficient $\alpha \in (0, 1)$ such that for all $\mu, \nu \in \mathcal{P}(\mathcal{X})$ (with $\mu \ll \nu$) and for all $t \geq 0$, we have

$$\mathbf{E}_\mu^\gamma \left[||\pi_{t+1}^\mu - \pi_{t+1}^\nu||_{\text{TV}} \right] \leq \alpha \mathbf{E}_\mu^\gamma \left[||\pi_t^\mu - \pi_t^\nu||_{\text{TV}} \right]. \tag{II.29}$$

We can make equivalent definitions for the filter process; 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 [53]. 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.

D. 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 $\hat{\gamma}$ to denote a policy that maps $\hat{Z}_t$ to U_t and we use $\hat{J}_\beta(\hat{z}_0, \hat{\gamma})$ to be some appropriately defined discounted cost under that policy. We have to be careful with these definitions as the approximate sequence $(\hat{Z}_t)_{t\geq 0}$ may not form a Markov chain, but we cover those technicalities later. Furthermore, when the approximation depends on some parameter N, we denote such a policy by $\hat{\gamma}_N$; when used in this way, we assume the policy is stationary, and thus the subscript N should not be confused with a

time index. In Section III, N determines the window length, and in Section V-A, it determines the "resolution" of quantization.

2) To denote *optimality* we use an asterisk symbol. For example, we denote $J^*_\beta(z_0) \coloneqq \min_\gamma J_\beta(z_0, \gamma)$, and we denote a policy achieving this minimum by γ^*.
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 $\hat{\gamma}$ and appropriately modify it to take Z_t as an input rather than $\hat{Z}_t$, we would call the resulting policy $\widetilde{\gamma}$.

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 $\hat{J}^*_\beta(\hat{z}_0)$ to all of $\mathcal{Z}$, not as the optimal discounted cost for some "extended" MDP.

III. Sliding Finite Window Approximation of the Belief MDP

Here we describe how to approximate π_t using a sliding finite window. The analysis in this section is inspired by [42], which used a similar construction to study sliding finite window policies for partially observed Markov decision processes (POMDPs). We note that although the proof methods are similar, our setup differs from [42] in two ways: 1) our problem is not strictly a POMDP as studied in [42], because the induced observation kernel changes with the choice of quantizer Q_t, 2) accordingly, our belief state is the predictor rather than the filter, and thus the Dobrushin coefficient terms we arrive at are different with a tailored analysis.

First, we must define a slightly different MDP than that defined in Proposition 2.8. 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:

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

$$W_t \coloneqq (\pi_{t-N}, I_t). \tag{III.2}$$

Note that we can compute π_t given w_t by applying the update equations in (II.15) N times. Denote this mapping by

$$\varphi : \mathcal{W} \to \mathcal{P}(\mathcal{X}) \tag{III.3}$$

$$W_t \mapsto \pi_t \tag{III.4}$$

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: This approach assumes that we start at time $t \geq N$; although for discounted MDPs 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 $\gamma \in \Gamma_{\mathrm{WS}}$. 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}).

A. The Sliding Finite Window Belief MDP

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 $t \geq 0$,

$$P(W_{t+1} \in \cdot | W_{[0,t]}, Q_{[0,t]}) = P(W_{t+1} \in \cdot | W_t, Q_t). \tag{III.5}$$

Proposition 3.1: 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) $c(w, Q) = \sum_{\mathcal{M}'} \min_{\hat{x}\in\hat{\mathcal{X}}} \sum_{\mathcal{X}} d(x, \hat{x}) O_Q(m'|x)\varphi(w)(x)$.

Note that the cost function is exactly the cost function we had in Proposition 2.8, by simply replacing $\pi = \varphi(w)$. Then an analog to Lemma 2.9 holds, and we indeed have that solving the MDP from Proposition 3.1 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 2.5 and the fact that $\pi_t = \varphi(W_t)$.

Proposition 3.2: 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, for all $w_0 \in \mathcal{W}$,

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

B. Sliding Finite Window Approximation

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

$$\hat{W}_t = (\zeta, I_t) \tag{III.6}$$

$$\hat{\pi}_t = \varphi(\hat{W}_t). \tag{III.7}$$

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

$$\pi_t(x) = P^\gamma_{\pi_{t-N}}(X_t = x | M'_{[t-N,t-1]}, Q_{[t-N,t-1]}) \tag{III.8}$$

$$\hat{\pi}_t(x) = P^\gamma_\zeta(X_t = x | M'_{[t-N,t-1]}, Q_{[t-N,t-1]}). \tag{III.9}$$

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 $\mathcal{W}_N \coloneqq \{\zeta\} \times (\mathcal{M}')^N \times \mathcal{Q}^N$. 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,

$$P_N(\hat{w}_1|\hat{w}, Q) \coloneqq P(\mathcal{P}(\mathcal{X}), i_1|\hat{w}, Q), \tag{III.10}$$

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

$$c_N(\hat{w}, Q) \coloneqq \sum_{\mathcal{M}'} \min_{\hat{x} \in \hat{\mathcal{X}}} \sum_{\mathcal{X}} d(x, \hat{x}) O_Q(m'|x) \hat{\pi}(x). \tag{III.11}$$

Then, our approximate MDP becomes $\text{MDP}_N = (\mathcal{W}_N, \mathcal{Q}, P_N, c_N)$. As in previous sections, denote the discounted cost for this MDP under a given policy $\hat{\gamma}_N$ (from $\mathcal{W}_N$ to $\mathcal{Q}$) by $\hat{J}_\beta(\hat{w}_0, \hat{\gamma}_N)$, and the optimal discounted cost by $\hat{J}^*_\beta(\hat{w}_0)$, with minimizing policy $\hat{\gamma}^*_N$. The relevant extensions (which are simply 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:

$$L^N_t \coloneqq \sup_{\gamma \in \Gamma_{\text{WS}}} \mathbf{E}^\gamma_{\pi_{t-N}} \left[||\pi_t - \hat{\pi}_t||_{\text{TV}} \right]. \tag{III.12}$$

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 $||d||_\infty \coloneqq \max_{x,\hat{x}} d(x, \hat{x})$, where we recall that d is the distortion measure for the zero-delay coding problem.

Theorem 3.3: For any $\gamma \in \Gamma_{\text{WS}}$ acting on N time steps to generate W_0 and any prior $\pi_{-N} \in \mathcal{P}(\mathcal{X})$, we have

$$\mathbf{E}^\gamma_{\pi_{-N}} \left[\left| \widetilde{J}^*_\beta(W_0) - J^*_\beta(W_0) \right| \right] \leq \frac{||d||_\infty}{1-\beta} \sum_{t=0}^{\infty} \beta^t L^N_t. \tag{III.13}$$

That is, we can bound the difference in the "extended" approximate optimal cost and the true optimal cost by a predictor-distance term.

Theorem 3.4: For any $\gamma \in \Gamma_{\text{WS}}$ acting on N time steps to generate w_0 and any prior $\pi_{-N} \in \mathcal{P}(\mathcal{X})$, we have

$$\mathbf{E}^\gamma_{\pi_{-N}} \left[\left| J_\beta(W_0, \widetilde{\gamma}^*_N) - J^*_\beta(W_0) \right| \right] \leq \frac{2||d||_\infty}{1-\beta} \sum_{t=0}^{\infty} \beta^t L^N_t. \tag{III.14}$$

That is, we can bound the sub-optimality of the "extended" approximate optimal policy $\widetilde{\gamma}^*_N$ by a predictor-distance term. Thus, if we can show that this distance goes to 0 as N grows, and in particular if we can determine a certain rate of convergence, we can get a similar rate of convergence to optimality for our policy $\widetilde{\gamma}^*_N$. The proofs for Theorems 3.3 and 3.4 are given in the Appendix.

C. 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 II-C). Indeed, the term within the supremum is exactly

$$\mathbf{E}^\gamma_\mu \left[||\pi^\mu_t - \pi^\nu_t||_{\text{TV}} \right] \tag{III.15}$$

when $\mu = \pi_{t-N}$ and $\nu = \zeta$, and under some $\gamma \in \Gamma_{\text{WS}}$. 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 [54], which uses joint contraction properties of the state and observation kernels to bound (III.15). 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

$$\delta(K) \coloneqq \inf \sum_{i=1}^{n} \min(K(B_i|x), K(B_i|y)), \tag{III.16}$$

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

$$K = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}. \tag{III.17}$$

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 [54, Theorem 3.6]. Note that in our case, the channel is not time-invariant, unlike in [54], but the analysis follows similarly. The proof is provided in the Appendix.

Theorem 3.5: For any $\gamma \in \Gamma_{\text{WS}}$ and for any $\mu \ll \nu$, we have

$$\mathbf{E}_\mu^\gamma \left[||\pi_{t+1}^\mu - \pi_{t+1}^\nu||_{\text{TV}} \right] \leq (1 - \delta(T))(2 - \tilde{\delta}(O)) \mathbf{E}_\mu^\gamma \left[||\pi_t^\mu - \pi_t^\nu||_{\text{TV}} \right], \tag{III.18}$$

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 3.6: Assume $\alpha \coloneqq (1 - \delta(T))(2 - \delta(O)) < 1$. Then for any $\gamma \in \Gamma_{\text{WS}}$ and for any $\mu \ll \nu$, we have

$$\mathbf{E}_\mu^\gamma \left[||\pi_{t+1}^\mu - \pi_{t+1}^\nu||_{\text{TV}} \right] \leq \alpha \mathbf{E}_\mu^\gamma \left[||\pi_t^\mu - \pi_t^\nu||_{\text{TV}} \right]. \tag{III.19}$$

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

$$\begin{aligned} L_t^N &= \sup_{\gamma \in \Gamma_{\text{ws}}} \mathbf{E}_{\pi_{t-N}}^\gamma \left[||\pi_t - \hat{\pi}_t||_{\text{TV}} \right] \\ &\leq \alpha^N ||\pi_{t-N} - \pi'||_{\text{TV}} \leq 2\alpha^N. \end{aligned} \tag{III.20}$$

The requirement $(1-\delta(T))(2-\delta(O)) < 1$ places conditions on the source dynamics and the channel. A universal condition to make the result applicable over all channels is obtained when one considers the special case where the channel is noiseless: in this case, we have $\delta(O) = 0$ and $\delta(T) > \frac{1}{2}$ is to hold. 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.

Combining Theorem 3.4 and the bound in (III.20), we obtain the following result.

Corollary 3.7: Assume $\alpha \coloneqq (1 - \delta(T))(2 - \delta(O)) < 1$. Then for any $\gamma \in \Gamma_{\text{WS}}$ which acts on N time steps to generate W_0 and any prior π_{-N}, we have

$$\mathbf{E}^{\gamma}_{\pi_{-N}} \left[\left| J_\beta(W_0, \widetilde{\gamma}^*_N) - J^*_\beta(W_0) \right| \right] \leq \frac{4||d||_\infty}{(1-\beta)^2} \alpha^N. \tag{III.21}$$

IV. Q-learning: Convergence to Near-Optimality

We now turn our attention to actually calculating an optimal policy for the approximate MDP in Section III, which will be near-optimal for the zero-delay coding problem for sufficiently large parameter N. That is, the remaining problem is to find for some large N the $\hat{\gamma}^*_N$ which minimizes

$$\mathbf{E}^{\hat{\gamma}^*_N}_{\pi_{-N}} \left[\sum_{t \geq 0} \beta^t c_N(\hat{w}_t, Q_t) \right]. \tag{IV.1}$$

We can then extend this policy in an appropriate manner to obtain a near-optimal policy $\widetilde{\gamma}^*$ for

$$\mathbf{E}^{\widetilde{\gamma}^*}_{\pi_0} \left[\sum_{t \geq 0} \beta^t c(\pi_t, Q_t) \right], \tag{IV.2}$$

which was equivalent to the discounted distortion in (II.6).

A tempting strategy from stochastic control theory would be the *value iteration* method (see [40, Lemma 4.2.8]), which iteratively computes

$$V_t(\hat{w}) = \min_{\mathcal{Q}} \left(c_N(\hat{w}, Q) + \beta \int_{\mathcal{W}_N} V_{t-1}(\hat{w}') P_N(d\hat{w}'|\hat{w}, Q) \right), \tag{IV.3}$$

for all $\hat{w} \in \mathcal{W}_N$ and with $V_0(\cdot) \equiv 0$. However, if we attempt to perform this method for the approximate zero-delay MDP, it becomes very analytically complicated. In particular, although we know the transition matrix T of the source $(X_t)_{t \geq 0}$, it is not clear how to compute P_N for the approximate MDP. This motivates the use of a reinforcement learning algorithm in which this transition kernel is not computed directly and the optimal policy is learned empirically. We study a version of Q-learning, originally from [55] for finite sets and generalized for weak Feller MDPs (along with rigorous convergence and near-optimality guarantees) in [56] and [42]. We first define the following update equations, where V_t and α_t are both functions from $\mathcal{Z} \times \mathcal{U}$ to $\mathbb{R}_+$, and $C_t \in \mathbb{R}_+$:

$$\begin{aligned} V_{t+1}(z,u) &= V_t(z,u) \quad \text{for all } (z,u) \neq (Z_t, U_t), \\ V_{t+1}(Z_t, U_t) &= (1 - \alpha_t(Z_t, U_t))\, V_t(Z_t, U_t) + \alpha_t(Z_t, U_t) \left[C_t + \beta \min_{v \in \mathcal{U}} V_t(Z_{t+1}, v) \right], \end{aligned} \tag{IV.4}$$

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

for some arbitrary V_0. Note that in the literature, Q is often used instead of V; here we reserve Q for quantizers. Our first algorithm is simply data collection - it samples X_t, Q_t, and M'_t to be used in our Q-learning iterations. Note that although the data collection and the Q-learning are presented as two separate algorithms for clarity, one could run the two concurrently: as we will see, all of the information needed to compute V_t is present at time t in the data collection.

Algorithm 1: Data Collection

Require: initial distribution π_0, transition kernel T, quantizer set $\mathcal{Q}$, channel kernel O

Sample $X_0 \sim \pi_0$
Choose Q_0 uniformly from $\mathcal{Q}$
$M_0 = Q_0(X_0)$
Sample $M'_0 \sim O(\cdot|M_0)$
for $t \geq 1$ **do**
  Sample $X_t \sim T(\cdot|X_{t-1})$
  Choose Q_t uniformly from $\mathcal{Q}$
  $M_t = Q_t(X_t)$
  Sample $M'_t \sim O(\cdot|M_t)$

This gives us a sequence $(Q_t, M'_t)_{t\geq 0}$. We then apply the iterations in (IV.4) with appropriate choices for Z_t, U_t and C_t. With an abuse of notation, let us shift the time index of the sampled data by N to obtain $(Q_t, M'_t)_{t\geq -N}$. Then apply the sliding finite window scheme from Section III to obtain $(\hat{W}_t)_{t\geq 0}$. Then we define Z_t, U_t and C_t as follows:

$$Z_t \coloneqq \hat{W}_t \tag{IV.5}$$

$$U_t \coloneqq Q_t \tag{IV.6}$$

$$C_t \coloneqq c_N(\hat{W}_t, Q_t). \tag{IV.7}$$

Remark: When this method is used in [42] for POMDPs, the realizations of the "true" cost $c(z,u)$ is used, since it is not assumed that we know the exact form of the cost function. However, in our case we wish to avoid computing π_t since it is not needed for the approximation, and computing c_N is straightforward via (III.11). Accordingly, we use the "approximate" cost $c_N(\hat{W}_t, Q_t)$.

Theorem 4.1: The sequence $(V_t)_{t\geq 0}$ converges almost surely to a limit V^*. Furthermore, consider the policy

$$\hat{\gamma}^*_N(\hat{w}) = \operatorname*{argmin}_{Q\in\mathcal{Q}} V^*(\hat{w}, Q), \tag{IV.8}$$

and let $\widetilde{\gamma}^*_N$ be the extension of $\hat{\gamma}^*_N$ to $\mathcal{W}$ by making it constant over the belief coordinate, as in Theorem 3.4. Then for any γ which acts on N time steps to generate W_0 and any prior π_{-N},

$$\mathbf{E}^{\gamma}_{\pi_{-N}}\left[J_\beta(W_0, \widetilde{\gamma}^*_N) - J^*_\beta(W_0)\right] \leq \frac{4||d||_\infty}{(1-\beta)^2}\alpha^N, \tag{IV.9}$$

where $\alpha = (1-\delta(T))(2-\delta(O))$.

The proof for Theorem 4.1 is given in the Appendix. Note that convergence relies on recent results in [57]; although convergence for finite MDPs is classical from [55], when applied to approximate MDPs (as we do here), convergence is not trivial since the underlying true MDP dynamics differ from that of the approximate MDP. Accordingly we must verify certain ergodic behavior of the MDP to apply the results in [57].

V. An Alternative Approximation Scheme and Comparison

A. Nearest Neighbor Approximation

Another approximation scheme for the zero-delay coding problem was proposed in [43], along with a Q-learning algorithm to compute the corresponding near-optimal policy. Although this scheme was only formally analyzed in the case of a noiseless channel, the results go through in the case of a noisy channel with some modifications. Accordingly, we present the noisy channel analogs of the results in [43] without proof, but we do provide a detailed comparison of the two schemes.

First, we approximate π_t using a nearest neighbor quantization scheme with a finite number of bins. This approximation can be done efficiently using [58, Algorithm 1], which approximates $\mathcal{P}(\mathcal{X})$ by the following finite set,

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

where $\sum_{i=1}^{N} k_i = N$. Note that in the literature, such distributions are also called types or empirical distributions [59]. We let $\hat{\pi}$ be the nearest neighbor (in Euclidean distance) of π in $\mathcal{P}_N(\mathcal{X})$, and we clearly have that $\max_\pi d(\pi, \hat{\pi}) \to 0$ as $N \to \infty$.

Similarly to Section III, one can then define a transition kernel and cost function over $\mathcal{P}_N(\mathcal{X})$. We omit the details, but they are essentially the averages of the true transition kernel $P(d\pi'|\pi, Q)$ and cost function $c(\pi, Q)$ over the nearest neighbor bins defined by (V.1), with respect to an appropriate reference measure. We denote the resulting approximate MDP by $\text{MDP}_N = (\mathcal{P}_N, \mathcal{Q}, P_N, c_N)$.

Corollary 5.1: [43, Theorem 1] Let $\hat{\gamma}_N^* \in \Gamma_{\text{ws}}$ 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 $\beta \in (0,1)$,

$$\lim_{N\to\infty} |J_\beta(\pi_0, \widetilde{\gamma}_N^*) - J_\beta^*(\pi_0)| = 0. \tag{V.2}$$

That is, $\widetilde{\gamma}_N^*$ is near-optimal for the zero-delay coding problem for sufficiently large N.

The analogous Q-learning result is as follows, using the sampled data obtained via Algorithm 1. Let $(\pi_t)_{t\geq 0}$ be the sequence of predictors resulting from the sampled data $(Q_t, M_t')_{t\geq 0}$ (with prior π_0). Recall that this can be computed recursively using the equations in (II.15). Apply the nearest neighbor approximation scheme onto $\mathcal{P}_N(\mathcal{X})$ to obtain $(\hat{\pi}_t)_{t\geq 0} \subset \mathcal{P}_N(\mathcal{X})$. Then, recalling the iterations in (IV.4), we define Z_t, U_t and C_t as follows:

$$Z_t \coloneqq \hat{\pi}_t \tag{V.3}$$

$$U_t \coloneqq Q_t \tag{V.4}$$

$$C_t \coloneqq c(\pi_t, Q_t) \quad \text{(recall (II.16))}. \tag{V.5}$$

Theorem 5.2: [43, Theorem 1] Assume that π_0 in Algorithm 1 is the unique invariant distribution of the source, ζ. Then V_t converges almost surely to a limit V^*. Furthermore, consider the policy

$$\hat{\gamma}_N^*(\hat{\pi}) = \operatorname*{argmin}_{Q \in \mathcal{Q}} V^*(\hat{\pi}, Q), \tag{V.6}$$

and let $\widetilde{\gamma}_N^*$ be the extension of $\hat{\gamma}_N^*$ to $\mathcal{P}(\mathcal{X})$ by making it constant over the belief quantization bins, as in Corollary 5.1. Then as $N \to \infty$,

$$J_\beta(\zeta, \widetilde{\gamma}_N^*) \to J_\beta^*(\zeta). \tag{V.7}$$

That is, $\widetilde{\gamma}_N^*$ is near-optimal for the discounted distortion zero-delay coding problem, provided that the source starts from its invariant distribution.

B. Reinforcement Learning Theoretic Comparison of the Two Schemes

In this section, we provide a detailed comparison noting explicit benefits and drawbacks of each of the two learning schemes to attain near-optimality. In particular, we emphasize the easier implementation and lower computational complexity of our proposed sliding window method, although there are some cases where the belief quantization method of [43] may be preferable.

- *Assumptions for convergence of Q-learning:* In Theorem 5.2, one can see that the convergence of V_t depends on the initial distribution used during learning (in particular, the π_0 in Algorithm 1 must be the source's unique invariant distribution ζ). Conversely, in Theorem 4.1, the convergence of V_t happens regardless of the initial distribution used during learning (that is, the π_0 in Algorithm 1 can be arbitrary).
- *Conditions on initialization during implementation:* For the belief quantization scheme, the policy from Q-learning is near-optimal when $\pi_0 = \zeta$. We note that in the special case of the noiseless channel, valid initializations also include $\pi_0 = \delta_x$, i.e., deterministically starting from any $x \in \mathcal{X}$. This can be shown using [43, Corollary 2] and by noting that in this special case the Dirac masses δ_x are recurrent. However, in the noisy channel case, this argument fails and we require $\pi_0 = \zeta$; conversely, for the sliding finite window scheme, the policy obtained in Theorem 4.1 is near-optimal for almost every initial window.
- *Filter stability assumptions:* Because the predictor stability required for the sliding finite window scheme is rather strong, the near-optimality of approximations was only established for those source-channel combinations satisfying the Dobrushin coefficient conditions. However, we note that the Dobrushin coefficent conditions are not necessary; any filter stability result which gives a bound on the loss term in Theorem 3.4 could be used. To this end, Hilbert metric approaches may also be relevant, for example those in [60], [61].
- *Computational complexity and need for Bayesian updates:* In the belief quantization scheme, 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. Conversely, 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.

	Convergence of Q-learning and near-optimality	Stability conditions not needed	Insensitive to initialization	Exponential convergence of performance	Bayesian update not needed	Lookup table implementation
Belief Quantization	✓	✓	✗	✗	✗	✓
Sliding Finite Window	✓	✗	✓	✓	✓	✓

TABLE I: Comparison of the two approximation schemes

- *Model knowledge and a data-driven approach:* 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 $\hat{X}_t$. Thus, the belief quantization approach requires knowledge, or at least a good estimate, of the underlying model (in particular, the initial distribution and the kernels T and O). Conversely, 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. For the decoder, if one has knowledge of the model, it can simply take the form of (II.8), as we did in our implementation. However, theoretically one could also apply learning at the decoder: treat the decoder as a map $\gamma^d : W_t, Q_t, M'_t \to \hat{\mathcal{X}}$ and add this as another dimension in the Q-table, so that one now learns a joint encoder-decoder policy. Under the same Dobrushin coefficient conditions as for the encoder, this will become near-optimal for large N. Of course, this comes at a significant memory and runtime cost for the Q-learning algorithm, so if one has model knowledge it is still beneficial to incorporate this at the decoder.
- *Rate of convergence to near-optimality:* For the sliding finite window result in Theorem 3.4, 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 for each initial finite window). Conversely, the belief quantization result in Corollary 5.1 is only asymptotic.
- *On quantization:* For the belief quantization result in Corollary 5.1, 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 [58, Algorithm 1] (which gives a uniform quantization). The sliding finite window can be seen as a non-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 I.

C. Implications for the Average Cost Problem

Under mild assumptions, it can be shown that as $\beta \to 1$, a near-optimal policy for the discounted cost problem becomes near-optimal for the average cost problem (see [41, Theorem 7.3.6]). It is straightforward to show that these assumptions hold for the zero-delay coding MDP; indeed, this same set of assumptions was shown to hold

in [27] and used to show the existence of optimal Walrand-Varaiya codes for the average cost problem. Then through [41, Theorem 7.3.6] and our previous results we obtain the following corollaries. Note that the dependence on β is only asymptotic; in practice we set $\beta = 0.9999$ and this seems to give policies very close to the optimum. For a fixed β, the dependence on ϵ is as given in Theorems 4.1 and 5.2 (exponential for sliding finite window, asymptotic for belief quantization).

Corollary 5.3: Let $\epsilon > 0$, let $\widetilde{\gamma}_N^*$ be the policy given in Theorem 4.1, and assume the conditions of Theorem 4.1 are met. Then there exists $\beta \in (0,1)$ and $N(\epsilon, \beta)$ sufficiently large such that for any γ which acts on the first N time steps to generate W_0 and any π_{-N},

$$\mathbf{E}^{\gamma}_{\pi_{-N}} \left[J(W_0, \widetilde{\gamma}_N^*) - J^*(W_0) \right] < \epsilon. \tag{V.8}$$

Corollary 5.4: Let $\epsilon > 0$, let $\widetilde{\gamma}_N^*$ be the policy given in Theorem 5.2. Then there exists $\beta \in (0,1)$ and $N(\epsilon, \beta)$ sufficiently large such that

$$J(\zeta, \widetilde{\gamma}_N^*) - J^*(\zeta) < \epsilon. \tag{V.9}$$

VI. Simulations

We now give some examples of zero-delay coding problems and simulate the performance of the policies resulting from Theorems 5.2 and 4.1. 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), so that the belief quantization scheme can be used. We observe that in each case with a known optimum, our Q-learning algorithm indeed reached this optimum (or very close to it), validating our theoretical results. In the case with an unknown optimum, we observe convergence rates which agree with the bounds in the theorems.

Remark: For the 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 $|\mathcal{P}_N(\mathcal{X})| = \binom{N+|\mathcal{X}|+1}{|\mathcal{X}|-1}$ [58], 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 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. Due to this, the schemes are somewhat difficult to compare on the same set of axes, since the N variable in each scheme leads to very different codebook sizes. Thus, when plotting against N on the x-axis we provide the results on separate sets of axes.

Also note that for a stopping criteria in the algorithms, any measure of the convergence of V_t would be suitable; in our implementation, we stop when the relative pointwise difference is sufficiently small. For example, $\max_{\hat{\pi},Q}((V_{t+1}(\hat{\pi}, Q) - V_t(\hat{\pi}, Q))/V_t(\hat{\pi}, Q) < \epsilon$ for some small $\epsilon > 0$.

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

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

$$T(\cdot|x) = \left(\tfrac{1}{4} \quad \tfrac{1}{8} \quad \tfrac{1}{8} \quad \tfrac{1}{16} \quad \tfrac{1}{16} \quad \tfrac{1}{16} \quad \tfrac{1}{4} \quad \tfrac{1}{16}\right). \tag{VI.1}$$

Note that in the i.i.d. case, we trivially have $\delta(T) = 1$, so Corollary 3.7 is applicable. Indeed, here we have that $\alpha = 0$, so that for all $N \geq 1$,

$$\mathbf{E}^{\gamma}_{\pi_{-N}}\left[\left|J_\beta(W_0, \widetilde{\gamma}^*) - J^*_\beta(W_0)\right|\right] = 0. \tag{VI.2}$$

That is, the optimal policy for the 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 $R \coloneqq \log_2 |\mathcal{M}|$. As expected by the above discussion, our algorithm matches with this quantizer in each case, which can be seen in Figure 2.

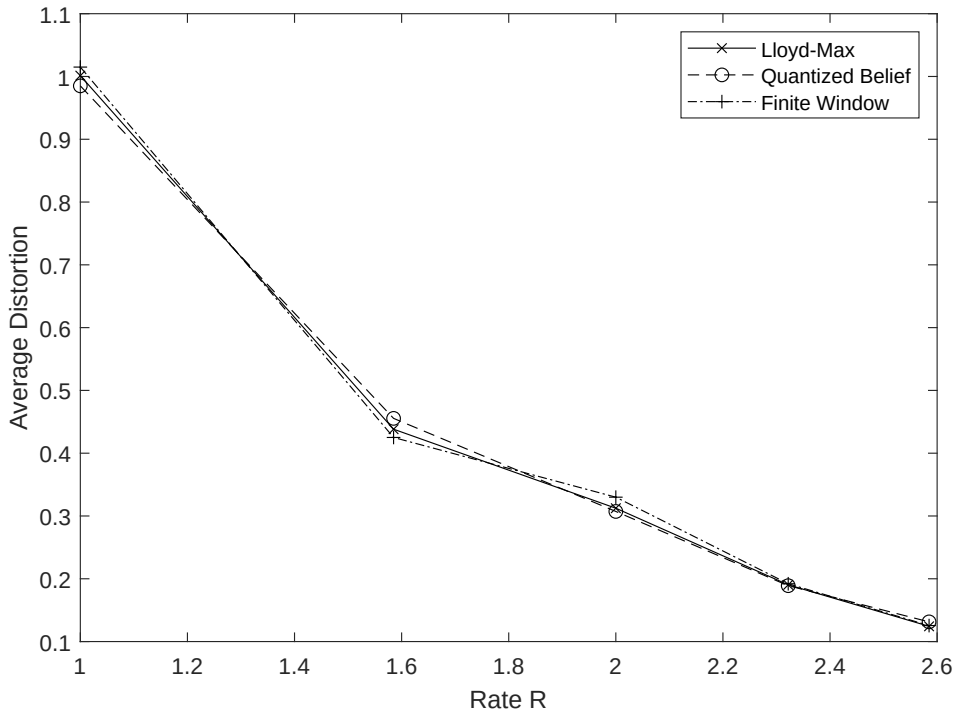


Fig. 2: Comparison with Lloyd-Max

B. Comparison with memoryless encoding

Consider now a Markov source with transition kernel given by

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

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

$$O = \begin{pmatrix} 0.94 & 0.02 & 0.02 & 0.02 \\ 0.02 & 0.94 & 0.02 & 0.02 \\ 0.02 & 0.02 & 0.94 & 0.02 \\ 0.02 & 0.02 & 0.02 & 0.94 \end{pmatrix}. \tag{VI.4}$$

We have that $\delta(T) = \frac{2}{3} > \frac{1}{2}$, so we can apply Corollary 3.7. In such a setup (where $\mathcal{X} = \mathcal{M}$ and the channel is symmetric), it was shown in [25] that "memoryless" encoding (that is, where $M_t = X_t$) is optimal. We compare our algorithms against such an encoding policy, shown in Figures 3 and 4, 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 finite window method. Accordingly, we present the schemes on different sets of axes. Note that the convergence of the finite window scheme to the optimum indeed appears exponential in N, as expected by Theorem 4.1.

C. 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

$$T = \begin{pmatrix} 0.2476 & 0.1527 & 0.0775 & 0.2219 & 0.2082 & 0.0920 \\ 0.0805 & 0.0247 & 0.0776 & 0.1290 & 0.3718 & 0.3164 \\ 0.1510 & 0.2335 & 0.2042 & 0.0107 & 0.1425 & 0.2580 \\ 0.0056 & 0.2252 & 0.2303 & 0.2173 & 0.1141 & 0.2076 \\ 0.1357 & 0.2685 & 0.0494 & 0.1981 & 0.2930 & 0.0553 \\ 0.2373 & 0.2795 & 0.0698 & 0.0399 & 0.1371 & 0.2363 \end{pmatrix}. \tag{VI.5}$$

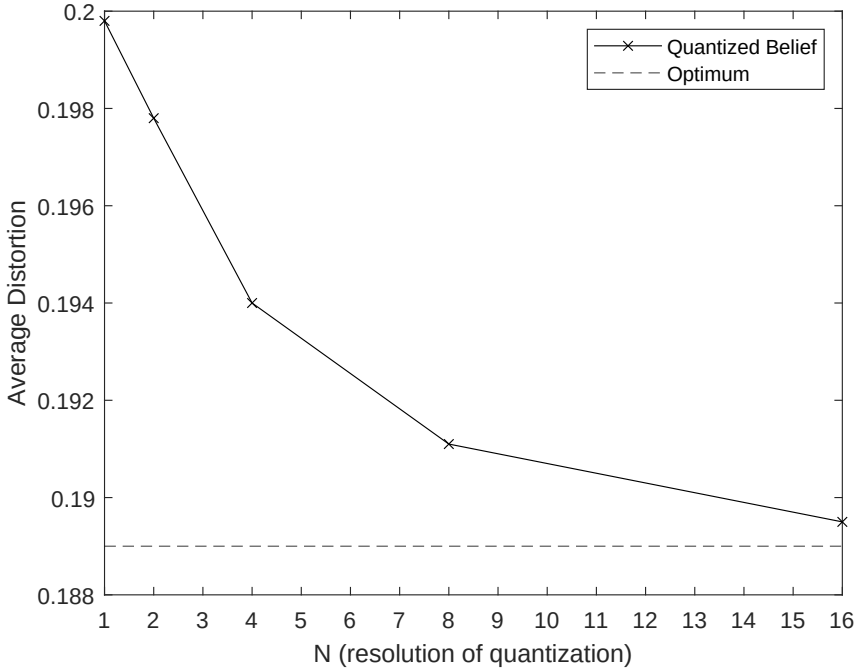


Fig. 3: Quantized belief scheme vs memoryless encoding

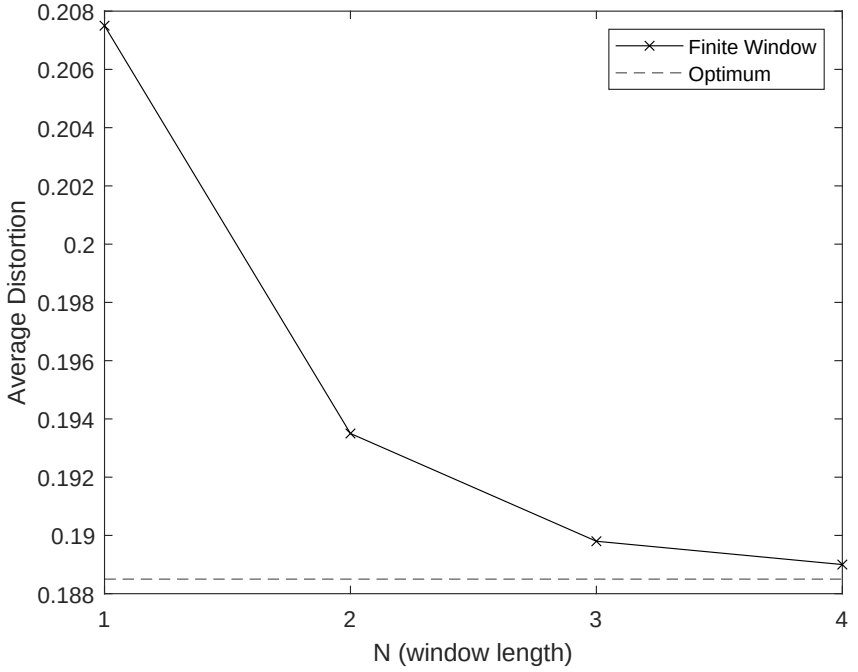


Fig. 4: Finite memory scheme vs memoryless encoding

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:

$$O = \begin{pmatrix} 0.96 & 0.02 & 0.02 \\ 0.02 & 0.96 & 0.02 \\ 0.02 & 0.02 & 0.96 \end{pmatrix}. \tag{VI.6}$$

It can be verified that the Dobrushin coefficient conditions of Theorem 4.1 are met for this setup. Figures 5 and 6 give the performance of both schemes for this problem.

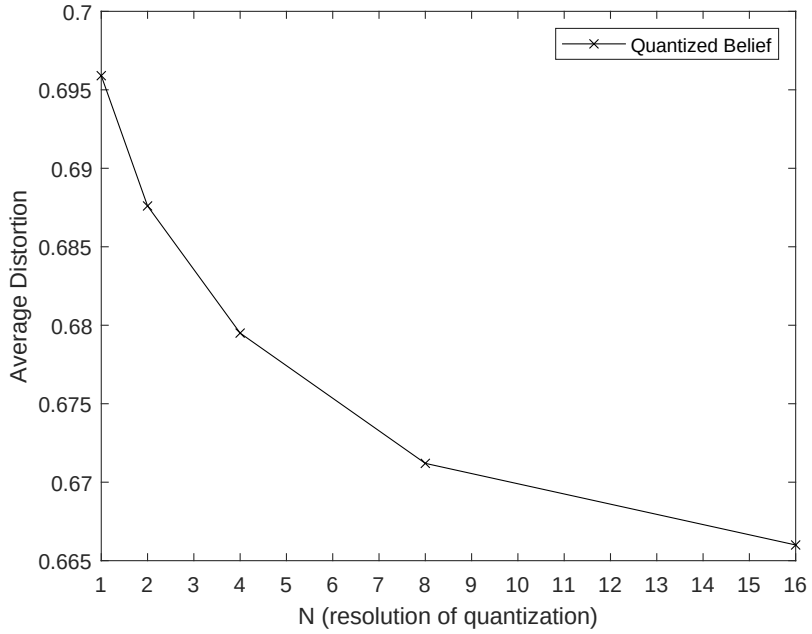


Fig. 5: Quantized belief scheme with unknown optimum

VII. Conclusion

We have provided two complementary approximation schemes for the zero-delay coding problem over a noisy channel with feedback, one in which the underlying belief space is quantized through a nearest-neighbor map, and the other in which the belief is approximated with a finite window of past observations. We then showed the convergence of a Q-learning algorithm to policies which are optimal for these approximations, yielding near-optimal policies for the zero-delay coding problem. Finally, we illustrated the convergence of our algorithm to the optimum through simulations. For future research, we wish to generalize our results to the case where our source-channel pair is continuous. We note that such a generalization should be mostly straightforward - under some recurrence assumptions on the source, one can still show the required regularity properties of the underlying MDP [39], and then most of the approximation and convergence results go through with minor technical changes. We also wish to find less stringent filter stability conditions for the finite window scheme than the Dobrushin coefficient ones given here; this may be possible via observability-type conditions, such as those in [62]. The controlled case, where one

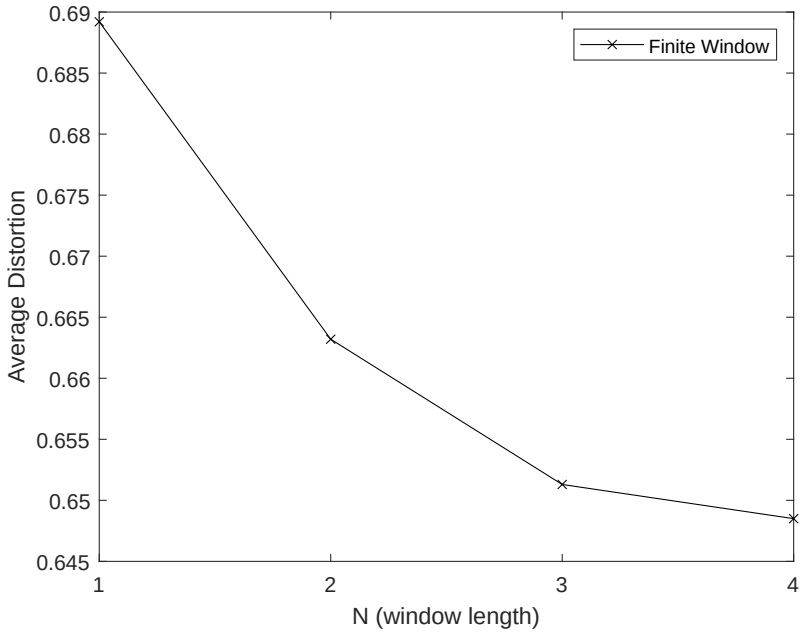


Fig. 6: Finite memory scheme with unknown optimum

communicates over a channel to a controller which affects some dynamical system, is also a promising research direction.

APPENDIX A
PROOF OF THEOREMS 3.3 AND 3.4

Lemma A.1: Consider the finite window approximation used in Section III, and 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

$$||P(M'_t \in \cdot | w_t, Q_t) - P(M'_t \in \cdot | \hat{w}_t, Q_t)||_{\text{TV}} \leq ||\pi_t - \hat{\pi}_t||_{\text{TV}}. \tag{A.1}$$

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

$$\left| \sum_{\mathcal{M}'} f(m'_t) P(m'_t | w_t, Q_t) - \sum_{\mathcal{M}'} f(m'_t) P(m'_t | \hat{w}_t, Q_t) \right| \tag{A.2}$$

$$= \left| \sum_{\mathcal{X}} \sum_{\mathcal{M}'} f(m'_t) P(m'_t | w_t, x_t, Q_t) P(x_t | w_t, Q_t) - \sum_{\mathcal{X}} \sum_{\mathcal{M}'} f(m'_t) P(m'_t | \hat{w}_t, x_t, Q_t) P(x_t | \hat{w}_t, Q_t) \right| \tag{A.3}$$

$$= \left| \sum_{\mathcal{X}} \sum_{\mathcal{M}'} f(m'_t) O_{Q_t}(m'_t | x_t) \pi_t(x_t) - \sum_{\mathcal{X}} \sum_{\mathcal{M}'} f(m'_t) O_{Q_t}(m'_t | x_t) \hat{\pi}_t(x_t) \right| \tag{A.4}$$

$$\leq ||\pi_t - \hat{\pi}_t||_{\text{TV}}, \tag{A.5}$$

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 $\gamma \in \Gamma_{\text{WS}}$. The last line follows from the fact that $g(x) \coloneqq \sum_{\mathcal{M}'} f(m')O_Q(m'|x)$ is upper bounded by 1. Taking the supremum over all such f yields the result. ■

Proof of Theorem 3.3

We provide a proof for the case when $N = 1$, but an analogous proof follows for $N > 1$. It can be shown (see [40, Theorem 4.2.3]) that the functions $J_\beta(w_t, \gamma)$ and $J^*_\beta(w_t)$ satisfy the following fixed-point equations:

$$J_\beta(w_t, \gamma) = c(w_t, \gamma(w_t)) + \beta \int_{\mathcal{W}} J_\beta(w_{t+1}, \gamma) P(dw_{t+1}|w_t, \gamma(w_t)) \tag{A.6}$$

$$J^*_\beta(w_t) = \min_{Q_t \in \mathcal{Q}} \Big(c(w_t, Q_t) + \beta \int_{\mathcal{W}} J^*_\beta(w_{t+1}) P(dw_{t+1}|w_t, Q_t) \Big), \tag{A.7}$$

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

$$J_\beta(w_t, \gamma) = c(w_t, \gamma(w_t)) + \beta \sum_{m'_t \in \mathcal{M}'} J_\beta((\pi_t, m'_t, \gamma(w_t)), \gamma) P(m'_t|w_t, \gamma(w_t)) \tag{A.8}$$

$$J^*_\beta(w_t) = \min_{Q_t \in \mathcal{Q}} \Big(c(w_t, Q_t) + \beta \sum_{m'_t \in \mathcal{M}'} J^*_\beta(\pi_t, m'_t, Q_t) P(m'_t|w_t, Q_t) \Big). \tag{A.9}$$

The functions $\hat{J}_\beta(\hat{w}_t, \hat{\gamma})$ and $\hat{J}^*_\beta(\hat{w}_t)$ satisfy equivalent fixed-point equations to (A.9), so that for $N = 1$,

$$\hat{J}_\beta(\hat{w}_t, \hat{\gamma}) = c_1(\hat{w}_t, \hat{\gamma}(\hat{w}_t)) + \beta \sum_{m'_t \in \mathcal{M}'} \hat{J}_\beta(\zeta, m'_t, \hat{\gamma}(\hat{w}_t)) P(m'_t|\hat{w}_t, \hat{\gamma}(\hat{w}_t)) \tag{A.10}$$

$$\hat{J}^*_\beta(\hat{w}_t) = \min_{Q_t \in \mathcal{Q}} \Big(c_1(\hat{w}_t, Q_t) + \beta \sum_{m'_t \in \mathcal{M}'} \hat{J}^*_\beta(\zeta, m'_t, Q_t) P(m'_t|\hat{w}_t, Q_t) \Big). \tag{A.11}$$

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

$$\beta \sum_{m'_0 \in \mathcal{M}'} \hat{J}^*_\beta(\zeta, m'_0, Q_0) P(m'_0|w_0, Q_0) = \beta \sum_{m'_0 \in \mathcal{M}'} \widetilde{J}^*_\beta(\pi_0, m'_0, Q_0) P(m'_0|w_0, Q_0). \tag{A.12}$$

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

$$\Big| \widetilde{J}^*_\beta(w_0) - J^*_\beta(w_0) \Big| \tag{A.13}$$

$$= \Big| \hat{J}^*_\beta(\hat{w}_0) - \beta \sum_{m'_0 \in \mathcal{M}'} \hat{J}^*_\beta(\zeta, m'_0, Q_0) P(m'_0|w_0, Q_0) \tag{A.14}$$

$$+ \beta \sum_{m'_0 \in \mathcal{M}'} \widetilde{J}^*_\beta(\pi_0, m'_0, Q_0) P(m'_0|w_0, Q_0) - J^*_\beta(w_0) \Big|. \tag{A.15}$$

Then applying the fixed-point equations (A.9) and (A.11) to the last and first terms respectively,

$$\Big| \widetilde{J}^*_\beta(w_0) - J^*_\beta(w_0) \Big|$$

$$\leq \max_{Q_0 \in \mathcal{Q}} |c_1(\hat{w}_0, Q_0) - c(w_0, Q_0)|$$

$$+\beta \max_{Q_0\in\mathcal{Q}} \left| \sum_{m_0'} \hat{J}^*_\beta(\zeta, m_0', Q_0) P(m_0'|\hat{w}_0, Q_0) - \sum_{m_0'} \hat{J}^*_\beta(\zeta, m_0', Q_0) P(m_0'|w_0, Q_0) \right|$$

$$+\beta \max_{Q_0\in\mathcal{Q}} \left| \sum_{m_0'} \widetilde{J}^*_\beta(\pi_0, m_0', Q_0) P(m_0'|w_0, Q_0) - \sum_{m_0'} J^*_\beta(\pi_0, m_0', Q_0) P(m_0'|w_0, Q_0) \right|.$$

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 $\gamma \in \Gamma_{\text{WS}}$, but we omit these in the expectation for notational simplicity. For the first term, by definition of c and c_N we have

$$\mathbf{E}\left[\left|c_1(\hat{W}_0, Q_0) - c(W_0, Q_0)\right|\right] \leq ||d||_\infty \mathbf{E}\left[||\hat{\pi}_0 - \pi_0||_{\text{TV}}\right] \leq ||d||_\infty L_0^1, \tag{A.16}$$

where $||d||_\infty = \max_{x,\hat{x}} d(x,\hat{x})$ and we recall the definition of L_t^N in (III.12); that is, $L_0^1 = \sup_{\gamma\in\Gamma_{\text{WS}}} \mathbf{E}\left[||\pi_t - \hat{\pi}_t||_{\text{TV}}\right]$. For the second term, we have

$$\mathbf{E}\left[\max_{Q_0\in\mathcal{Q}} \left| \sum_{m_0'} \hat{J}^*_\beta(\zeta, m_0', Q_0) P(m_0'|\hat{W}_0, Q_0) - \sum_{m_0'} \hat{J}^*_\beta(\zeta, m_0', Q_0) P(m_0'|W_0, Q_0) \right|\right] \tag{A.17}$$

$$\leq ||\hat{J}^*_\beta||_\infty \mathbf{E}\left[\max_{Q_0\in\mathcal{Q}} ||P(m_0'|\hat{W}_0, Q_0) - P(m_0'|W_0, Q_0)||_{\text{TV}}\right] \tag{A.18}$$

$$\leq ||\hat{J}^*_\beta||_\infty L_0^1, \tag{A.19}$$

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 A.1. Finally, since both sums in the last term are over the same measure $P(m_0'|W_0, Q_0)$, we have

$$\mathbf{E}\left[\max_{Q_0\in\mathcal{Q}} \left| \sum_{m_0'} \widetilde{J}^*_\beta(\pi_0, m_0', Q_0) P(m_0'|W_0, Q_0) - \sum_{m_0'} J^*_\beta(\pi_0, m_0', Q_0) P(m_0'|W_0, Q_0) \right|\right] \tag{A.20}$$

$$\leq \sup_{\gamma'\in\Gamma_{\text{WS}}} \mathbf{E}\left[\left|\widetilde{J}^*_\beta(W_1) - J^*_\beta(W_1)\right|\right], \tag{A.21}$$

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

$$\mathbf{E}\left[\left|\widetilde{J}^*_\beta(W_0) - J^*_\beta(W_0)\right|\right] \tag{A.22}$$

$$\leq (||d||_\infty + \beta||\hat{J}^*_\beta||_\infty) L_0^1 + \beta \sup_{\gamma'\in\Gamma_{\text{WS}}} \mathbf{E}\left[\left|\widetilde{J}^*_\beta(W_1) - J^*_\beta(W_1)\right|\right] \tag{A.23}$$

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

$$\mathbf{E}\left[\left|\widetilde{J}^*_\beta(W_0) - J^*_\beta(W_0)\right|\right] \leq \frac{||d||_\infty}{1-\beta} \sum_{t=0}^{\infty} \beta^t L_t^1. \tag{A.24}$$

■

Proof of Theorem 3.4

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 (A.8),

$$J_\beta(w_0, \widetilde{\gamma}_1^*) = c(w_0, Q_0^*) + \beta \sum_{m_0' \in \mathcal{M}'} J_\beta((\pi_0, m_0', Q_0^*), \widetilde{\gamma}_1^*) P(m_0'|w_0, Q_0^*) \tag{A.25}$$

and using (A.11) and the fact that $\widetilde{J}_\beta^*(w_0) = \hat{J}_\beta^*(\hat{w_0})$,

$$\widetilde{J}_\beta^*(w_0) = c_1(\hat{w}_0, Q_0^*) + \beta \sum_{m_0' \in \mathcal{M}'} \widetilde{J}_\beta^*(\pi_0, m_0', Q_0^*) P(m_0'|\hat{w}_0, Q_0^*). \tag{A.26}$$

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

$$\sum_{m_0' \in \mathcal{M}'} \widetilde{J}_\beta^*(w_1) P(m_0'|w_0, Q_0^*), \tag{A.27}$$

to obtain

$$\begin{aligned}
&\left| J_\beta(w_0, \widetilde{\gamma}_1^*) - \widetilde{J}_\beta^*(w_0) \right| && \text{(A.28)}\\
\leq\ & |c(w_0, Q_0^*) - c(\hat{w}_0, Q_0^*)| && \text{(A.29)}\\
&+ \beta \sum_{m_0' \in \mathcal{M}'} \left| \widetilde{J}_\beta^*(w_1) P(m_0'|\hat{w}_0, Q_0^*) - \widetilde{J}_\beta^*(w_1) P(m_0'|w_0, Q_0^*) \right| && \text{(A.30)}\\
&+ \beta \sum_{m_0' \in \mathcal{M}'} \left| \widetilde{J}_\beta^*(w_1) - J_\beta(w_1, \widetilde{\gamma}_1^*) \right| P(m_0'|w_0, Q_0^*). && \text{(A.31)}
\end{aligned}$$

Thus, using (A.16) and Lemma A.1,

$$\begin{aligned}
&\mathbf{E}\left[\left| J_\beta(W_0, \widetilde{\gamma}_1^*) - \widetilde{J}_\beta^*(W_0) \right| \right] && \text{(A.32)}\\
\leq\ & ||d||_\infty L_0^1 && \text{(A.33)}\\
&+ \beta ||\widetilde{J}_\beta^*||_\infty \sup_{\gamma \in \Gamma_{\text{ws}}} \mathbf{E}\left[||P(m_0'|\hat{W}_0, Q_0^*) - P(m_0'|W_0, Q_0^*)||_{\text{TV}} \right] && \text{(A.34)}\\
&+ \beta \sup_{\gamma \in \Gamma_{\text{ws}}} \mathbf{E}\left[\left| \widetilde{J}_\beta^*(W_1) - J_\beta(W_1, \widetilde{\gamma}_1^*) \right| \right] && \text{(A.35)}\\
\leq\ & (||d||_\infty + \beta ||\widetilde{J}_\beta^*||_\infty) L_0^1 + \beta \sup_{\gamma'} \mathbf{E}\left[\left| J_\beta(W_1, \widetilde{\gamma}_1^*) - \widetilde{J}_\beta^*(W_1) \right| \right]. && \text{(A.36)}
\end{aligned}$$

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

$$\mathbf{E}\left[\left| J_\beta^*(W_0, \widetilde{\gamma}_1^*) - J_\beta^*(W_0) \right| \right] \leq \frac{||d||_\infty}{1-\beta} \sum_{t=0}^{\infty} \beta^t L_t^1. \tag{A.37}$$

Finally, we have

$$\begin{aligned}
&\mathbf{E}\left[\left| J_\beta^*(W_0, \widetilde{\gamma}_1^*) - J_\beta^*(W_0) \right| \right] && \text{(A.38)}\\
\leq\ & \mathbf{E}\left[\left| J_\beta^*(W_0, \widetilde{\gamma}_1^*) - \widetilde{J}_\beta^*(W_0) \right| \right] + \mathbf{E}\left[\left| \widetilde{J}_\beta^*(W_0) - J_\beta^*(W_0) \right| \right] && \text{(A.39)}\\
\leq\ & \frac{2||d||_\infty}{1-\beta} \sum_{t=0}^{\infty} \beta^t L_t^1, && \text{(A.40)}
\end{aligned}$$

where the final inequality follows from (A.37) and Theorem 3.3. ■

APPENDIX B
PROOF OF THEOREM 3.5

Lemma B.1: The following holds:

$$\sum_{\mathcal{X}}\sum_{\mathcal{M}'}||\overline{\pi}_t^\mu-\overline{\pi}_t^\nu||_{\text{TV}}O_{Q_t}(m'|x)\pi_t^\mu(x)\leq(2-\tilde{\delta}(O))||\pi_t^\mu-\pi_t^\nu||_{\text{TV}}, \tag{B.1}$$

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

Proof: The following argument is from [54, 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 (II.15), 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 $Q_t=Q$,

$$\left|\sum_{\mathcal{X}}f(x)\overline{\pi}_t^\mu(x)-\sum_{\mathcal{X}}f(x)\overline{\pi}_t^\nu(x)\right| \tag{B.2}$$

$$=\left|\sum_{\mathcal{X}}f(x)\frac{g_Q(x,m')\pi_t^\mu(x)}{N^\mu(m',Q)}-\sum_{\mathcal{X}}f(x)\frac{g_Q(x,m')\pi_t^\nu(x)}{N^\nu(m',Q)}\right| \tag{B.3}$$

$$\leq\left|\sum_{\mathcal{X}}f(x)\frac{g_Q(x,m')\pi_t^\mu(x)}{N^\mu(m',Q)}-\sum_{\mathcal{X}}f(x)\frac{g_Q(x,m')\pi_t^\nu(x)}{N^\mu(m',Q)}\right| \tag{B.4}$$

$$+\left|\sum_{\mathcal{X}}f(x)\frac{g_Q(x,m')\pi_t^\nu(x)}{N^\mu(m',Q)}-\sum_{\mathcal{X}}f(x)\frac{g_Q(x,m')\pi_t^\nu(x)}{N^\nu(m',Q)}\right| \tag{B.5}$$

$$=\frac{1}{N^\mu(m',Q)}\left|\sum_{\mathcal{X}}f(x)g_Q(x,m')\pi_t^\mu(x)-\sum_{\mathcal{X}}f(x)g_Q(x,m')\pi_t^\nu(x)\right| \tag{B.6}$$

$$+\left|\frac{1}{N^\mu(m',Q)}-\frac{1}{N^\nu(m',Q)}\right|\cdot\left|\sum_{\mathcal{X}}f(x)g_Q(x,m')\pi_t^\nu(x)\right| \tag{B.7}$$

$$\leq\frac{1}{N^\mu(m',Q)}\sum_{\mathcal{X}}g_Q(x,m')|\pi_t^\mu-\pi_t^\nu|(x) \tag{B.8}$$

$$+\left|\frac{1}{N^\mu(m',Q)}-\frac{1}{N^\nu(m',Q)}\right|N^\nu(m',Q) \tag{B.9}$$

$$\leq\frac{1}{N^\mu(m',Q)}\left(\sum_{\mathcal{X}}g_Q(x,m')|\pi_t^\mu-\pi_t^\nu|(x)+|N^\mu(m',Q)-N^\nu(m',Q)|\right), \tag{B.10}$$

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 $\sum_{\mathcal{X}}|\pi_t^\mu-\pi_t^\nu|(x)=||\pi_t^\mu-\pi_t^\nu||_{\text{TV}}$. Taking the supremum over all f gives

$$||\overline{\pi}_t^\mu-\overline{\pi}_t^\nu||_{\text{TV}}\leq\frac{1}{N^\mu(m',Q)}\left(\sum_{\mathcal{X}}g_Q(x,m')|\pi_t^\mu-\pi_t^\nu|(x)+|N^\mu(m',Q)-N^\nu(m',Q)|\right). \tag{B.11}$$

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

$$\sum_{\mathcal{X}}\sum_{\mathcal{M}'}||\overline{\pi}_t^\mu-\overline{\pi}_t^\nu||_{\text{TV}}O_{Q_t}(m'|x)\pi_t^\mu(x) \tag{B.12}$$

$$=\sum_{\mathcal{X}}\sum_{\mathcal{M}'}||\overline{\pi}_t^\mu-\overline{\pi}_t^\nu||_{\text{TV}}g_{Q_t}(x,m')\psi(m')\pi_t^\mu(x) \tag{B.13}$$

$$=\sum_{\mathcal{M}'}||\overline{\pi}_t^\mu-\overline{\pi}_t^\nu||_{\text{TV}}\left(\sum_{\mathcal{X}}g_{Q_t}(x,m')\pi_t^\mu(x)\right)\psi(m') \tag{B.14}$$

$$= \sum_{\mathcal{M}'} ||\overline{\pi}_t^\mu - \overline{\pi}_t^\nu||_{\text{TV}} N^\mu(m', Q_t)\psi(m') \tag{B.15}$$

$$\leq \sum_{\mathcal{M}'} \Big(\sum_{\mathcal{X}} g_{Q_t}(x, m')|\pi_t^\mu - \pi_t^\nu|(x) + |N^\mu(m', Q_t) - N^\nu(m', Q_t)| \Big) \psi(m') \tag{B.16}$$

$$\leq \sum_{\mathcal{X}} \left(\sum_{\mathcal{M}'} g_{Q_t}(x, m')\psi(m') \right) |\pi_t^\mu - \pi_t^\nu|(x) \tag{B.17}$$

$$+ \sum_{\mathcal{M}'} \left| \sum_{\mathcal{X}} g_{Q_t}(x, m')(\pi_t^\mu - \pi_t^\nu)(x) \right| \psi(m') \tag{B.18}$$

$$\leq ||\pi_t^\mu - \pi_t^\nu||_{\text{TV}} + \sum_{\mathcal{M}'} |O_{Q_t}(\pi_t^\mu) - O_{Q_t}(\pi_t^\nu)|\,(m') \tag{B.19}$$

$$= ||\pi_t^\mu - \pi_t^\nu||_{\text{TV}} + ||O_{Q_t}(\pi_t^\mu) - O_{Q_t}(\pi_t^\nu)||_{\text{TV}}, \tag{B.20}$$

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 [63] that the Dobrushin coefficient acts as a contraction coefficient for kernel operators under total variation. In particular

$$||O_{Q_t}(\pi_t^\mu) - O_{Q_t}(\pi_t^\nu)||_{\text{TV}} \leq (1 - \delta(O_{Q_t}))||\pi_t^\mu - \pi_t^\nu||_{\text{TV}}. \tag{B.21}$$

Thus,

$$\sum_{\mathcal{X}} \sum_{\mathcal{M}'} ||\overline{\pi}_t^\mu - \overline{\pi}_t^\nu||_{\text{TV}} O_{Q_t}(m'|x)\pi_t^\mu(x) \tag{B.22}$$

$$\leq (2 - \delta(O_{Q_t}))||\pi_t^\mu - \pi_t^\nu||_{\text{TV}} \tag{B.23}$$

$$\leq (2 - \tilde{\delta}(O))||\pi_t^\mu - \pi_t^\nu||_{\text{TV}}, \tag{B.24}$$

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

Proof of Theorem 3.5

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 $\gamma \in \Gamma_{\text{WS}}$, $Q_{[0,t]}$ are deterministic given μ and $M'_{[0,t-1]}$. Thus,

$$\mathbf{E}_\mu^\gamma \left[||\overline{\pi}_t^\mu - \overline{\pi}_t^\nu||_{\text{TV}} \right] \tag{B.25}$$

$$= \sum_{(\mathcal{M}')^{t+1}} ||\overline{\pi}_t^\mu - \overline{\pi}_t^\nu||_{\text{TV}} P_\mu^\gamma(m'_{[0,t]}) \tag{B.26}$$

$$= \sum_{(\mathcal{M}')^t} \sum_{\mathcal{X}} \sum_{\mathcal{M}'} ||\overline{\pi}_t^\mu - \overline{\pi}_t^\nu||_{\text{TV}} P_\mu^\gamma(m'_t|x_t, m'_{[0,t-1]}) P_\mu^\gamma(x_t|m'_{[0,t-1]}) P_\mu^\gamma(m'_{[0,t-1]}) \tag{B.27}$$

$$= \sum_{(\mathcal{M}')^t} \sum_{\mathcal{X}} \sum_{\mathcal{M}'} ||\overline{\pi}_t^\mu - \overline{\pi}_t^\nu||_{\text{TV}} O_{Q_t}(m'_t|x_t)\pi_t^\mu(x) P_\mu^\gamma(m'_{[0,t-1]}) \tag{B.28}$$

$$\leq (2 - \tilde{\delta}(O)) \sum_{(\mathcal{M}')^t} ||\pi_t^\mu - \pi_t^\nu||_{\text{TV}} P_\mu^\gamma(m'_{[0,t-1]}) \tag{B.29}$$

$$= (2 - \tilde{\delta}(O)) \mathbf{E}_\mu^\gamma \left[||\pi_t^\mu - \pi_t^\nu||_{\text{TV}} \right], \tag{B.30}$$

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 B.1. Finally, using the Dobrushin contraction property for kernels (as noted in the derivation of (B.21)), we have

$$\mathbf{E}_\mu^\gamma \left[||\pi_{t+1}^\mu - \pi_{t+1}^\nu||_{\text{TV}}\right] \leq (1-\delta(T))\mathbf{E}_\mu^\gamma \left[||\overline{\pi}_t^\mu - \overline{\pi}_t^\mu||_{\text{TV}}\right] \tag{B.31}$$

$$\leq (1-\delta(T))(2-\tilde{\delta}(O))\mathbf{E}_\mu^\gamma \left[||\pi_t^\mu - \pi_t^\nu||_{\text{TV}}\right]. \tag{B.32}$$

■

APPENDIX C
PROOF OF THEOREM 4.1

We will show certain ergodic behavior (in particular, [57, Assumption 2.1]) and then invoke [57, Theorem 2.1]. Indeed, we have trivially that

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

so that [57, Assumption 2.2 (ii)] holds. Additionally, by positive Harris recurrence of $(X_t)_{t\geq 0}$, the marginals on $P(X_t \in \cdot)$ converge to ζ, so we have for any f that

$$\frac{\sum_{k=0}^t f(\hat{w}_{k+1})\mathbf{1}(\hat{w}_k = \hat{w}, Q_k = Q)}{\sum_{k=0}^t \mathbf{1}(\hat{w}_k = \hat{w}, Q_k = Q)} \to \int f(\hat{w}_1) P_N(\hat{w}_1|\hat{w}, Q) \tag{C.2}$$

almost surely as $t \to \infty$, where P_N is from (III.10). Thus [57, Assumption 2.1 (iii)] holds.

Finally, note that not every $\hat{w} \in \mathcal{W}_N$ has positive probability of being visited (certain sequences of channel outputs and quantizers are impossible depending on the source and channel), so we restrict ourselves to only those with positive probability. Then [57, Assumption 2.1 (i)] holds for (almost every) $\hat{w}$, so we apply [57, Theorem 2.1] to obtain that almost surely, as $N \to \infty$,

$$V_t(\hat{w}, Q) \to V^*(\hat{w}, Q), \tag{C.3}$$

where

$$V^*(\hat{w}, Q) = c_N(\hat{w}, Q) + \beta \sum_{\hat{w}_1 \in \mathcal{W}_N} \min_Q V^*(\hat{w}_1) P_N(\hat{w}_1|\hat{w}, Q), \tag{C.4}$$

where P_N and c_N are from (III.10) and (III.11). Taking the minimum over $\mathcal{Q}$ gives us the classic Bellman optimality equation, and hence $\hat{J}_\beta(\hat{w}, \hat{\gamma}_N^*) = \hat{J}_\beta^*(\hat{w})$ for almost every $\hat{w} \in \mathcal{W}_N$. The result follows by applying Corollary 3.7. ■

REFERENCES

[1] A. Badr, A. Khisti, W.-T. Tan, and J. Apostolopoulos, "Streaming codes for channels with burst and isolated erasures," in *2013 Proceedings IEEE INFOCOM*. IEEE, 2013, pp. 2850–2858.

[2] M. Rudow and K. Rashmi, "Streaming codes for variable-size messages," *IEEE Transactions on Information Theory*, vol. 68, no. 9, pp. 5823–5849, 2022.

[3] A. Khisti and S. Draper, "The streaming-DMT of fading channels," *IEEE Transactions on Information Theory*, vol. 60, no. 11, pp. 7058–7072, 2014.

[4] A. Badr, P. Patil, A. Khisti, W.-T. Tan, and J. Apostolopoulos, "Layered constructions for low-delay streaming codes," *IEEE Transactions on Information Theory*, vol. 63, no. 1, pp. 111–141, 2017.

[5] A. Khina, V. Kostina, A. Khisti, and B. Hassibi, "Tracking and control of Gauss-Markov processes over packet-drop channels with acknowledgments," *IEEE Transactions on Control of Network Systems*, vol. 6, no. 2, pp. 549–560, 2018.

[6] M. Gastpar, B. Rimoldi, and M. Vetterli, "To code, or not to code: Lossy source-channel communication revisited," *IEEE Transactions on Information Theory*, vol. 49, pp. 1147–1158, May 2003.

[7] M. Gastpar, "Uncoded transmission is exactly optimal for a simple Gaussian sensor network," *IEEE Transactions on Information Theory*, vol. 54, pp. 5247–5251, 2008.

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

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

[19] J. Chakravorty and A. Mahajan, "Fundamental limits of remote estimation of autoregressive Markov processes under communication constraints," *IEEE Transactions on Automatic Control*, vol. 62, no. 3, pp. 1109–1124, 2017.

[20] N. Guo and V. Kostina, "Optimal causal rate-constrained sampling for a class of continuous Markov processes," *IEEE Transactions on Information Theory*, vol. 67, no. 12, pp. 7876–7890, 2021.

[21] T. Soleymani, J. S. Baras, and K. H. Johansson, "State estimation over delayed and lossy channels: An encoder–decoder synthesis," *IEEE Transactions on Automatic Control*, vol. 69, no. 3, pp. 1568–1583, 2024.

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

[23] 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.

[24] T. Linder, *Learning-theoretic methods in vector quantization*. Springer, Wien, New York, 2002.

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

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

[27] 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.

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

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

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

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] ——, "Feedback capacity of Ising channels with large alphabet via reinforcement learning," *IEEE Transactions on Information Theory*, vol. 68, no. 9, pp. 5637–5656, 2022.

[37] 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.

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

[39] 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.

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

[41] S. Yüksel. (2023) Optimization and Control of Stochastic Systems. [Online]. Available: https://mast.queensu.ca/ yuksel/LectureNotesOnStochasticOptControl.pdf

[42] 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.

[43] 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.

[44] F. Pollara, R. J. McEliece, and K. Abdel-Ghaffar, "Finite-state codes," *IEEE Transactions on Information Theory*, vol. 34, no. 5, pp. 1083–1089, 1988.

[45] J. C. Kieffer and J. G. Dunham, "On a type of stochastic stability for a class of encoding schemes," *IEEE Transactions on Information Theory*, vol. 29, pp. 793–797, November 1983.

[46] G. Ungerboeck, "Channel coding with multilevel/phase signals," *IEEE Transactions on Information Theory*, vol. 28, no. 1, pp. 55–67, 1982.

[47] A. Calderbank and J. Mazo, "A new description of trellis codes," *IEEE Transactions on Information Theory*, vol. 30, no. 6, pp. 784–791, 1984.

[48] R. Johannesson and K. S. Zigangirov, *Fundamentals of Convolutional Coding*. John Wiley & Sons, 2015.

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

[50] 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.

[51] R. van Handel, "Stochastic calculus, filtering, and stochastic control," *Course notes., URL: http://www. princeton. edu/˜ rvan/acm217/ACM217.pdf*, 2007.

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

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

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

[55] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, pp. 279–292, 1992.

[56] 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.

[57] 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.

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

[59] I. Csiszar, "The method of types [information theory]," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2505–2523, 1998.

[60] F. L. Gland and N. Oudjane, "Stability and uniform approximation of nonlinear filters using the Hilbert metric and application to particle filters," *The Annals of Applied Probability*, vol. 14, no. 1, pp. 144–187, 2004.

[61] S. Y. Yunus Emre Demirci, Ali Devran Kara, "Refined bounds on near optimality finite window policies in POMDPs and their reinforcement learning," *arXiv:2409.04351*, 2024.

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

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