Near-Optimality of Finite-Memory Codes and Reinforcement Learning for Zero-Delay Coding of Markov Sources

Liam Cregg, Fady Alajaji, Serdar Yüksel

***Abstract*— We study the problem of zero-delay coding of a Markov source over a noisy channel with feedback. Building and generalizing prior work, we first formulate the problem as a Markov decision process (MDP) where the state is a probability measure valued predictor along with a finite memory of channel outputs and quantizers. We then approximate this state by marginalizing over all possible predictors, so that our policies only use the finite-memory term to encode the source. Under an appropriate notion of predictor stability, we show that such policies are near-optimal for the zero-delay coding problem as the memory length increases. We also give sufficient conditions for predictor stability to hold, and present a reinforcement learning algorithm and establish its convergence to compute near-optimal finite-memory policies. These theoretical results are supported by simulations.**

I. Introduction

The zero-delay coding problem involves compressing and transmitting an information source 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 and real-time sensor networks. Totally noiseless feedback (which we assume in this paper) is often assumed in applications where the party using the backward channel has sufficiently high power that they can render the feedback essentially noiseless (compared to the forward channel). For example a phone sending information to a cell tower, or a communication problem where one channel is wireless and another is wired. Note that the zero-delay restriction means that classical Shannon-theoretic methods, which are generally asymptotic in nature, are not viable.

Within the information theory literature, the problem of joint source-channel coding (JSCC) with feedback is well-studied. Although a classic result due to Shannon [1] states that feedback does not increase the capacity of a memoryless channel, it has been shown that feedback does improve the delay-distortion tradeoff (also called the error exponent) and can simplify the design of coding schemes [2], [3]. Deep learning methods for JSCC have been proposed in e.g., [4], [5], although these often lack formal convergence and performance guarantees, and are not strictly zero-delay.

There has been success in studying this problem using stochastic control techniques. In particular, [6], [7] consider finite alphabets and finite time horizons and show optimality of restricted classes of policies. Similar optimality and existence results are presented for infinite time horizons in [8]–[10]. The continuous-alphabet, infinite-horizon case is studied in [11], [12]. However, these stochastic control techniques often utilize a state that is probability measure-valued (such a state is often called the "predictor" in the literature). This state space is computationally difficult to work with, and thus actually obtaining effective coding schemes for a given zero-delay coding problem is still an open problem.

There has been much recent work on learning-theoretic methods in stochastic control for very general state and action spaces. In particular, approximation techniques leading to near-optimal solutions are well-suited to setups such as the zero-delay coding problem, where exact methods are intractable. Some of these techniques use quantization schemes to approximate the underlying state space; for example, refer to [13], [14] and related reinforcement learning algorithms in [15], [16]. While these quantization-based methods can be used to find near-optimal solutions, the quantization process introduces additional computational complexity, both during learning and implementation.

In order to avoid the computational overhead of quantization, we will instead build on recent results from [17] which uses a finite history of past observations and control actions. The work in [17] also provides a Q-learning solution to obtain such near-optimal policies, which we will adapt to our setup. This approach relies on filter stability, which is a measure of how quickly a process forgets its prior as it collects observations. Our results will use the related question of predictor stability. Under predictor stability, this finite history acts as a good approximation of the true predictor, leading to explicit performance bounds on the resulting policy for the zero-delay coding problem. Our approach complements [18]. For a detailed review of filter stability methods in the control-free case, see [19]. However, we will see that in the zero-delay coding problem, the observations depend not only on the noisy channel but also on how we encode the information at the receiver. Most results in the literature assume that the observation kernel is time-invariant, which is not applicable here; so we will prove and use generalized results for when this kernel is control-dependent. In particular, we will generalize recent results from [20], which uses joint properties of the state and observation kernels to obtain exponential filter stability, as well as discuss recent observability-type conditions from [21], [22].

Additional work related to finite-memory policies can be found in [23], [24]. However, [23] does not provide a performance bound based on window length, and [24]

This work was supported by the Natural Sciences and Engineering Research Council of Canada. The authors are with the Department of Mathematics and Statistics, Queen's University, Kingston ON, Canada (email: {liam.cregg, fa, yuksel}@queensu.ca).

still utilizes a nearest-neighbour quantization map. Furthermore, [24] only provides near-optimality of the discounted-cost criteria for discount factors bounded away from one; in the zero-delay coding problem, we are generally interested in taking discount factors very close to one in order to obtain approximations to the average-cost problem.

Contributions: In this paper, we prove near-optimality of finite-memory coding policies for the zero-delay coding problem over a noisy channel with feedback, and explicitly bound the performance of such policies in terms of a predictor stability term. We generalize some existing predictor stability results to the case where the observation kernel is control-dependent, and apply these results to some example zero-delay coding problems. We also discuss a Q-learning algorithm that converges to such a near-optimal policy, and provide supporting simulation results. Due to space constraints, proofs and further literature review and simulations are included in [25].

II. Notation and Preliminaries

Let $\mathbb{X}$ be a finite set and let our information source be a Markov process $(x_t)_{t\geq 0}$ taking values in $\mathbb{X}$. Let $T(x_{t+1}|x_t)$ be its transition kernel, which we assume is irreducible and aperiodic, and thus admits a unique invariant measure. Let $x_0 \sim \pi_0$ (we will also call π_0 the prior). Let $\mathcal{M}$ and $\mathcal{M}'$ be the input and output alphabets of the noisy channel, which we assume are finite, and let $(q_t)_{t\geq 0}$ and $(q'_t)_{t\geq 0}$ be the respective processes. We will denote the channel kernel by $O(q'_t|q_t)$, which gives the probability of the channel output being q'_t given its input is q_t. Finally, let $\hat{\mathbb{X}}$ be some finite set of reconstruction values, and let $(\hat{x}_t)_{t\geq 0}$ take values in $\hat{\mathbb{X}}$. Throughout, given some sequence $(a_t)_{t\geq 0}$, we will use the notation $a_{[n,k]} \coloneqq (a_n, a_{n+1}, \dots, a_k)$ for $n \leq k$. Also, although everything is finite here, for ease of notation as well as providing easier generalizations to uncountable spaces, we will often write sums as integrals over the appropriate measure. For example, rather than writing $\sum_{x'} T(x'|x)$ we may write $\int_{\mathbb{X}} T(dx'|x)$.

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

$$\begin{aligned} \gamma^e_t : \mathbb{X}^{t+1} \times \mathcal{M}^t \times (\mathcal{M}')^t \to \mathcal{M} \quad \gamma^d_t : (\mathcal{M}')^{t+1} \to \hat{\mathbb{X}} \\ (x_{[0,t]}, q_{[0,t-1]}, q'_{[0,t-1]}) \mapsto q_t \qquad\qquad q'_{[0,t]} \mapsto \hat{x}_t. \end{aligned}$$

We consider the discounted-cost criterion for the zero-delay coding problem, which is to find encoder and decoder policies such that the following is minimized:

$$J_\beta(\pi_0, \gamma^e, \gamma^d) \coloneqq \mathbf{E}^{\gamma^e,\gamma^d}_{\pi_0}\left[\sum_{t=0}^{\infty} \beta^t d(x_t, \hat{x}_t)\right], \qquad (1)$$

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

Note that the discounted-cost criterion is not the standard objective for the zero-delay coding problem; we are usually concerned with the average-cost criterion. However, it can be shown that as $\beta \to 1$, a policy that is near-optimal for the discounted-cost problem in this setting is also near-optimal for the average-cost problem (e.g., [26, Theorem 7.3.6]. Thus, all of the results in this paper also hold for the average-cost criterion by taking β sufficiently close to 1. Indeed, in our simulations we will take $\beta = 0.99$ to obtain an approximation of the average-cost criterion.

Since all our sets are finite, there always exists an optimal decoder policy for a given encoder policy; so without loss of optimality, we can search only for an optimal encoder policy and assume that it is paired with an optimal decoder policy. We will denote such a joint encoder-decoder policy by $\gamma \coloneqq (\gamma^e, \gamma^{d*})$, where γ^{d*} is an optimal decoder policy given γ^e. We denote the infimum of (1) by

$$J^*_\beta(\pi_0) \coloneqq \inf_{\gamma} J_\beta(\pi_0, \gamma).$$

For fixed $(x_{[0,t-1]}, q_{[0,t-1]}, q'_{[0,t-1]})$, consider the function $\gamma^e(\cdot, x_{[0,t-1]}, q_{[0,t-1]}, q'_{[0,t-1]}) : \mathbb{X} \to \mathcal{M}$. Such a function (i.e., a mapping from $\mathbb{X}$ to $\mathcal{M}$) is called a *quantizer*. We denote the set of all such quantizers by $\mathcal{Q}$. Thus we can view an encoder policy γ^e as selecting a quantizer $Q_t \in \mathcal{Q}$ based on the information $(x_{[0,t-1]}, q_{[0,t-1]}, q'_{[0,t-1]})$, then generating the channel input q_t as $Q_t(x_t)$.

Recall that we used $O(q'_t|q_t)$ to denote our channel transition kernel. Let $O_{Q_t}(q'_t|x_t)$ denote the observation kernel induced by a quantizer $Q_t \in \mathcal{Q}$; that is, $O_{Q_t}(q'_t|x_t) = O(q'_t|Q_t(x_t))$. Denoting the set of probability measures on a set A by $\mathcal{P}(A)$, let $\psi \in \mathcal{P}(\mathcal{M}')$ be such that $O_Q(\cdot|x) \ll \psi$ for all $x \in \mathbb{X}, Q \in \mathcal{Q}$, where we use "$\ll$" to denote absolute continuity (i.e., $\psi(B) = 0 \implies O_Q(B|x) = 0$ for any Borel $B \subset \mathcal{M}'$). Since $\mathcal{M}'$ is finite in our setup, we could 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 $g_Q(x, q') \coloneqq \frac{dO_Q}{d\psi}(x, q')$ be the Radon-Nikodym derivative of O_Q with respect to ψ.

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

$$\begin{aligned} \pi_t(x_t) &= P^{\gamma}_{\pi_0}(x_t|q'_{[0,t-1]}) \\ \overline{\pi}_t(x_t) &= P^{\gamma}_{\pi_0}(x_t|q'_{[0,t]}), \end{aligned}$$

recalling that $X_0 \sim \pi_0$. We have dropped the γ for notational simplicity, but it should be noted that such measures are policy-dependent. With a slight abuse of notation, we also let T act as an operator on probability measures as follows:

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

Then given π_0, the above measures can be computed in a recursive manner as follows (see e.g., [27, Proposition

3.2.5]).

$$\overline{\pi}_t(dx) = \frac{g_{Q_t}(x, q'_t)\pi_t(dx)}{\int_{\mathbb{X}} g_{Q_t}(x, q'_t)\pi_t(dx)},$$
$$\pi_{t+1} = T(\overline{\pi}_t). \tag{2}$$

We will denote $N(q'_t, Q_t) := \int_{\mathbb{X}} g_{Q_t}(x, q'_t)\pi_t(dx)$. Note that $N(q'_t, Q_t)$ is non-zero $P^{\gamma}_{\pi_0}$ a.s. Thus inside of $P^{\gamma}_{\pi_0}$ expectations we will assume $N(q'_t, Q_t)$ is non-zero.

Using the above update equations, one can compute π_t given $(q'_{[0,t-1]}, Q_{[0,t-1]})$, so that policies of the form $Q_t = \gamma_t(\pi_t)$ are valid. We call such policies *Walrand-Varaiya* policies. If such a policy does not depend on t (i.e., $\gamma_t = \overline{\gamma}$ for some $\overline{\gamma}$ and all $t \geq 0$), then we call this policy *stationary*. The following is a key result, originally from Walrand and Varaiya [6] for a finite time horizon and extended to the infinite-horizon case in [8].

Proposition 1: [8, Proposition 2] For any $\beta \in (0, 1)$, there exists a stationary Walrand-Varaiya type policy γ^* that solves the discounted cost problem, i.e., one that satisfies $J_\beta(\pi_0, \gamma^*) = J^*_\beta(\pi_0)$ for all $\pi_0 \in \mathcal{P}(\mathbb{X})$.

Although a very useful existence result, the above does not lend itself well to numerical methods since $\mathcal{P}(\mathbb{X})$ is uncountable and the transition kernel for $(\pi_t)_{t\geq 0}$ is complex. In the next section, we will split π_t into a past prior and a finite memory of observations, then show that under certain regularity conditions we can replace the past prior with a constant. This yields a finite state space which is more amenable to numerical methods (in particular, reinforcement learning).

III. Finite-Memory-Belief Construction

We now construct our finite-memory-belief policies. The analysis in this section is inspired by [17], which used a similar construction to study finite-memory policies for Partially Observed Markov Decision Problems (POMDPs). We fix some $N \in \mathbb{Z}_+$, which we call the *memory* and for $t \geq N$ define

$$I^N_t = (q'_{[t-N,t-1]}, Q_{[t-N,t-1]})$$
$$z^N_t = (\pi_{t-N}, I^N_t).$$

Note that we can compute π_t given z^N_t by applying the update equations in (2) N times. Denote this mapping by

$$\varphi : \mathcal{Z} \to \mathcal{P}(\mathbb{X})$$
$$z^N_t \mapsto \pi_t$$

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

We will call policies of the form $Q_t = \gamma_t(z^N_t)$ *finite-memory-belief* policies (with memory N). Similarly, if it does not depend on t, we call it stationary.

Proposition 2: For any $\beta \in (0, 1)$ and $N \in \mathbb{Z}_+$, there exists a stationary finite-memory-belief policy that solves the discounted-cost problem, i.e., one that satisfies $J_\beta(\pi_N, \gamma^*) = J^*_\beta(\pi_N)$ for all $\pi_N \in \mathcal{P}(\mathbb{X})$.

Remark: Note that z^N_t is only defined for $t \geq N$. If the first N steps are significant (i.e, β is small) then a finite-memory-belief policy may not be near-optimal for a process starting at $t = 0$. However, recall that for the zero-delay coding problem we are generally interested in the average-cost cost criterion (by taking β close to 1). Accordingly, we assume that the distortion from the first N steps is negligible.

Properties of the Finite-Memory-Belief Construction

It can be shown that the process $(z^N_t)_{t\geq N}$ is controlled Markov (similarly to [17]), with control $(Q_t)_{t\geq N}$. That is, for all $t \geq N$,

$$P(z^N_{t+1}|z^N_{[N,t]}, Q_{[N,t]}) = P(z^N_{t+1}|z^N_t, Q_t) =: \eta(z^N_{t+1}|z^N_t, Q_t),$$

We introduce the following cost function $c : \mathcal{Z} \times \mathcal{Q} \to \mathbb{R}_+$,

$$c(z^N_t, Q_t) = \int_{\mathcal{M}'} \left(\min_{\hat{x}\in\hat{\mathbb{X}}} \int_{\mathbb{X}} d(x, \hat{x}) g_{Q_t}(x, q')\varphi(z^N_t)(dx) \right) \psi(dq').$$

Lemma 1: Given z^N_t and Q_t, $c(z^N_t, Q_t)$ is the expected distortion when the optimal decoder is used.

In fact, the above is simply a more general form for the cost function found in [8], [28]. Then by the assumption that we use the optimal decoder for a given encoder (and that the first N steps are negligible), we can write (1) as

$$J_\beta(z^N_N, \gamma) = \mathbf{E}^\gamma \left[\sum_{t=N}^{\infty} \beta^t c(z^N_t, Q_t) \right],$$

where γ is a finite-memory-belief policy, and we define $J^*_\beta(z^N_N) := \min_\gamma J_\beta(z^N_N, \gamma)$.

It can be shown (e.g., [29, Theorem 4.2.3]) that these functions satisfy the following fixed-point equations:

$$J_\beta(z^N_t, \gamma) = c(z^N_t, \gamma(z^N_t)) + \beta \int_{\mathcal{Z}} J_\beta(z^N_{t+1}, \gamma)\eta(dz^N_{t+1}|z^N_t, \gamma(z^N_t))$$
$$J^*_\beta(z^N_t) = \min_{Q_t\in\mathcal{Q}} \Big(c(z^N_t, Q_t) + \beta \int_{\mathcal{Z}} J^*_\beta(z^N_{t+1})\eta(dz^N_{t+1}|z^N_t, Q_t) \Big),$$

for all $z^N_t \in \mathcal{Z}$ and finite-memory-belief policy γ. Note that although the integral is over $\mathcal{Z}$, which is uncountable, we can only reach finitely many elements from a given z^N_t since the observation space $\mathcal{M}'$ is finite. In particular, when $N = 1$ and $t \geq 1$, we can write $z^1_t = (\pi_{t-1}, q'_{t-1}, Q_{t-1})$ and $z^1_{t+1} = (\varphi(z^1_t), q'_t, Q_t)$, so the above becomes

$$J_\beta(z^1_t, \gamma) = c(z^1_t, \gamma(z^1_t)) + \beta \sum_{q'_t\in\mathcal{M}'} J_\beta((\varphi(z^1_t), q'_t, \gamma(z^1_t)), \gamma)P(q'_t|z^1_t, \gamma(z^1_t)) \tag{3}$$

$$J^*_\beta(z^1_t) = \min_{Q_t\in\mathcal{Q}} \Big(c(z^1_t, Q_t) + \beta \sum_{q'_t\in\mathcal{M}'} J^*_\beta(\varphi(z^1_t), q'_t, Q_t)P(q'_t|z^1_t, Q_t) \Big). \tag{4}$$

One can rewrite these equations in a similar way for $N > 1$, but for simplicity we will usually study the case of $N = 1$.

IV. Finite-Memory Construction

The above representation is not particularly useful, as it still requires one to compute π_{t-N}. So we use the following approximation of z_t^N: fix some $\hat{\pi} \in \mathcal{P}(\mathbb{X})$ and let $\hat{z}_t^N = (\hat{\pi}, I_t^N)$. That is, $\hat{z}_t^N$ uses $\hat{\pi}$ as the predictor at time $t-N$, regardless of the true predictor. We can similarly apply φ to $\hat{z}_t^N$ to obtain an "incorrect" predictor at time t. The key idea is that under predictor stability the correct predictor $\varphi(z_t^N)$ and the incorrect predictor $\varphi(\hat{z}_t^N)$ will be close for large enough N, since the predictor will be insensitive to the prior. For technical reasons regarding predictor stability, we assume that $\hat{\pi}$ has full support over $\mathbb{X}$.

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

Properties of the Finite-Memory Construction

Consider the following transition kernel given by taking the marginal of η over $\mathcal{P}(\mathbb{X})$,

$$\hat{\eta}(\hat{z}_{t+1}^N | \hat{z}_t^N, Q_t) \coloneqq \eta(\mathcal{P}(\mathbb{X}), I_{t+1}^N | \hat{z}_t^N, Q_t).$$

Also consider the cost function $c(\hat{z}_t^N, Q_t)$ and the resulting value function, which we denote by

$$\hat{J}_\beta(\hat{z}_N^N, \hat{\gamma}) \coloneqq \mathbf{E}^{\hat{\gamma}} \left[\sum_{t=N}^{\infty} \beta^t c(\hat{z}_t^N, Q_t) \right],$$

where the policy $\hat{\gamma}$ maps $\hat{z}_t^N$ to Q_t, and we denote $\hat{J}_\beta^*(\hat{z}_N^N) \coloneqq \min_{\hat{\gamma}} \hat{J}_\beta(\hat{z}_N^N, \hat{\gamma})$. Note that a minimizing policy, which we denote by $\hat{\gamma}^*$, exists trivially in this case since $\hat{z}_t^N$ and Q_t can take only finitely many values. The above functions satisfy equivalent fixed-point equations to (4), so that for $N = 1$,

$$\begin{aligned} \hat{J}_\beta(\hat{z}_t^1, \hat{\gamma}) &= c(\hat{z}_t^1, \hat{\gamma}(\hat{z}_t^1)) \\ &\quad + \beta \sum_{q_t' \in \mathcal{M}'} \hat{J}_\beta(\hat{\pi}, q_t', \hat{\gamma}(\hat{z}_t^1)) P(q_t' | \hat{z}_t^1, \hat{\gamma}(\hat{z}_t^1)) \end{aligned}$$

$$\begin{aligned} \hat{J}_\beta^*(\hat{z}_t^1) = \min_{Q_t \in \mathcal{Q}} \Big(& c(\hat{z}_t^1, Q_t) \\ & + \beta \sum_{q_t' \in \mathcal{M}'} \hat{J}_\beta^*(\hat{\pi}, q_t', Q_t) P(q_t' | \hat{z}_t^1, Q_t) \Big). \end{aligned} \quad (5)$$

Note that we can extend $\hat{J}_\beta^*$ and $\hat{\gamma}^*$ to all of $\mathcal{Z}$ by making them constant over $\mathcal{P}(\mathbb{X})$. We denote these extensions by $\widetilde{J}_\beta^*$ and $\widetilde{\gamma}^*$, so that

$$\begin{aligned} \widetilde{J}_\beta^*(z^N) &= \widetilde{J}_\beta^*(\pi, I^N) \coloneqq \hat{J}_\beta^*(\hat{\pi}, I^N) \\ \widetilde{\gamma}^*(z^N) &= \widetilde{\gamma}^*(\pi, I^N) \coloneqq \hat{\gamma}^*(\hat{\pi}, I^N), \end{aligned} \quad (6)$$

for all $z^N = (\pi, I^N) \in \mathcal{Z}$.

We are interested in the value of $\left| J_\beta(z_N^N, \widetilde{\gamma}^*) - J_\beta^*(z_N^N) \right|$; that is, the loss in performance when we apply the optimal policy from the finite-memory representation to the finite-memory-belief representation (with the appropriate extension).

Remark: The process $(\hat{z}_t^N)_{t \geq N}$ is in general not controlled Markov. That is, in reality $\hat{z}_t^N$ does not have a transition kernel given by $\hat{\eta}$; in this section we have constructed an artificial MDP with this transition kernel. However, the long-term ergodic behaviour of $\hat{z}_t^N$ will coincide with the fixed-point equations in (5) through the Q-learning approach to be covered in Section VII. That is, we will use $\hat{z}_t^N$ to find optimal policies for this artificial MDP, even though $\hat{z}_t^N$ itself does not form an MDP.

V. Loss in Performance Due to Approximation

We define the *total variation distance* between two probability measures μ, ν as

$$||\mu - \nu||_{TV} \coloneqq \sup_{||f||_\infty \leq 1} \left| \int f d\mu - \int f d\nu \right|,$$

where f is measurable. Using our above definition of $c(z^N, Q)$, we have

$$\left| c(z^N, Q) - c(\hat{z}^N, Q) \right| \leq ||d||_\infty ||\varphi(z^N) - \varphi(\hat{z}^N)||_{TV}. \quad (7)$$

Lemma 2: For any z_t^N and $\hat{z}_t^N$, under any finite-memory-belief policy, we have

$$||P(q_t' | z_t^N, Q_t) - P(q_t' | \hat{z}_t^N, Q_t)||_{TV} \leq ||\varphi(z_t^N) - \varphi(\hat{z}_t^N)||_{TV}.$$

In the following, we bound the difference in the optimal expected costs between the finite-memory-belief representation and the finite-memory representation. We will use this to obtain a performance bound later. Note that we consider an expectation with respect to z_N^N (that is, over the first N steps). The following is a loss term: for $t \geq N$, let

$$L_t^N \coloneqq \sup_\gamma \mathbf{E}_{\pi_0}^\gamma \left[||\varphi(z_t^N) - \varphi(\hat{z}_t^N)||_{TV} \right], \quad (8)$$

where the supremum is over all policies that generate I_{N-1}^N (i.e., that act on the first N time steps).

Theorem 1: Let γ be some policy that generates I_{N-1}^N, then

$$\mathbf{E}_{\pi_0}^\gamma \left[\left| \widetilde{J}_\beta^*(z_N^N) - J_\beta^*(z_N^N) \right| \right] \leq \frac{||d||_\infty}{1-\beta} \sum_{t=N}^{\infty} \beta^t L_t^N.$$

Theorem 2: Let γ be some policy that generates I_{N-1}^N, and let $\widetilde{\gamma}^*$ be the optimal policy for the finite-memory representation extended to $\mathcal{Z}$ as in (6). Then,

$$\mathbf{E}_{\pi_0}^\gamma \left[\left| J_\beta(z_N^N, \widetilde{\gamma}^*) - J_\beta^*(z_N^N) \right| \right] \leq \frac{2||d||_\infty}{1-\beta} \sum_{t=N}^{\infty} \beta^t L_t^N.$$

VI. Predictor Stability Conditions

The loss term in the previous theorems,

$$L_t^N \coloneqq \sup_\gamma \mathbf{E}_{\pi_0}^\gamma \left[||\varphi(z_t^N) - \varphi(\hat{z}_t^N)||_{TV} \right],$$

is the expected total variation distance between the predictors at time t, given that the predictors at time $t-N$ are given by π_{t-N} and $\hat{\pi}$, respectively. This is related to the question of *predictor stability*, which we now review.

Note that the update equations in (2) are sensitive to the choice of π_0. We modify the previous notation for π_t and $\overline{\pi}_t$ to indicate this dependence: let $(\pi_t^\mu)_{t\geq 0}$ and $(\pi_t^\nu)_{t\geq 0}$ be the predictors resulting from the same sequence $(q_t', Q_t)_{t\geq 0}$, but with different initial measures $\pi_0^\mu = \mu$ and $\pi_0^\nu = \nu$, where $\mu \ll \nu$. We equivalently define $\overline{\pi}_t^\mu$ and $N^\mu(q_t', Q_t)$.

We wish to study the behaviour of

$$\mathbf{E}_\mu^\gamma \left[||\pi_t^\mu - \pi_t^\nu||_{TV}\right] \tag{9}$$

as $t \to \infty$, under some policy γ. Bounding this term over all γ will give us a bound on L_t^N above by taking $\mu = \pi_{t-N}$ and, for example, $\nu = \hat{\pi}$ to be uniform over $\mathbb{X}$ (so that the absolute continuity condition is satisfied).

Dobrushin Coefficient Conditions

The following results are inspired by the analysis in [20], which uses joint contraction properties of the state and observation kernels to bound (9). In particular, we will study the following type of stability:

Definition 1: A predictor process is called *exponentially stable* in total variation if for some $\alpha < 1$.

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

First we introduce some notation. For standard Borel spaces A_1, A_2 and some kernel $K : A_1 \to \mathcal{P}(A_2)$, we define the Dobrushin coefficient as

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

where the infimum is over $x, y \in A_1$ and all partitions $\{B_i\}_{i=1}^n$ of 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. Then we have a counterpart of [20, Theorem 3.6] in the case where the channel is not time-invariant.

Theorem 3: For any finite-memory-belief policy γ and for any $\mu \ll \nu$, we have

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

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

Corollary 1: For any policy γ and for any $\mu \ll \nu$,

$$\mathbf{E}_\mu^\gamma \left[||\pi_t^\mu - \pi_t^\nu||_{TV}\right] \leq \alpha^t ||\mu - \nu||_{TV},$$

where $\alpha \coloneqq (1-\delta(T))(2-\tilde{\delta}(O))$. Thus if $\alpha < 1$, then the predictor is exponentially stable in total variation with coefficient α. Also, if $\delta(T) > \frac{1}{2}$, then the predictor is exponentially stable regardless of the channel.

Note that for a given quantizer Q, the kernel $O_Q(q'|x) = O(q'|Q(x))$ only contains rows from the kernel O, so that $\delta(O) \leq \delta(O_Q)$ for all Q. Thus we obtain the following.

Corollary 2: For any policy γ and for any $\mu \ll \nu$,

$$\mathbf{E}_\mu^\gamma \left[||\pi_t^\mu - \pi_t^\nu||_{TV}\right] \leq \alpha^t ||\mu - \nu||_{TV},$$

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

Note that, in many applications of the zero-delay quantization problem, the requirement that $(1-\delta(T))(2-\delta(O)) < 1$ is too strong. In particular, in the special case where the channel is noiseless, we will always have that $\delta(O) = 0$. Thus we can only use Corollary 2 if $\delta(T) > \frac{1}{2}$. This is not surprising given the nature of Dobrushin-type conditions; the more similar the conditional measures $O_Q(dq'|x)$ and $O_Q(dq'|y)$ are for different $x, y \in \mathbb{X}$, the closer the Dobrushin coefficient is to 1. Therefore, such Dobrushin-type conditions prioritize *uninformative* kernels. Conversely, the goal of the zero-delay coding problem is to use quantizers that create *informative* kernels. Nevertheless, the above conditions give an easy-to-verify condition for predictor stability.

Applying Corollary 2 to the L_t^N term, we have

$$\begin{aligned}L_t^N &= \sup_\gamma \mathbf{E}_{\pi_0}^\gamma \left[||\varphi(z_t^N) - \varphi(\hat{z}_t^N)||_{TV}\right] \\ &\leq \alpha^N ||\pi_{t-N} - \hat{\pi}||_{TV}.\end{aligned}$$

VII. Reinforcement Learning for Finite-Memory Policies

In order to obtain such a finite-memory policy, we propose using a variation of the well-known Q-learning algorithm [30]. Q-learning is a reinforcement learning algorithm in which realizations of state, action, and cost are collected and used to update Q-factors whose limits, under certain assumptions, can be used to obtain an optimal policy. This algorithm cannot be applied directly to the process $(\hat{z}_t^N, Q_t)_{t\geq 0}$ as it is not Markov. Due to the above issues, we use the more general version of the Q-learning algorithm proposed in [31, Theorem 2.1] (see also [17, Theorem 4.1]). In this algorithm, a policy γ' is applied and realizations of $(\hat{z}_t^N, Q_t, c(\hat{z}_t^N, Q_t))_{t\geq N}$ are collected. Consider the sequence $(\mathsf{Q}_t)_{t\geq N}$ where $\mathsf{Q}_t : \hat{\mathcal{Z}} \times \mathcal{Q} \to \mathbb{R}_+$.

For some initial Q_N, we compute this sequence using

$$\begin{aligned}\mathsf{Q}_{t+1}(\hat{z}, Q) = (1-\alpha_t(\hat{z}, Q))\mathsf{Q}_t(\hat{z}, Q) \\ + \alpha_t(\hat{z}, Q)\left(c(\hat{z}, Q) + \beta \min_v \mathsf{Q}_t(\hat{z}_{t+1}^N, v)\right)\end{aligned}, \tag{10}$$

where $\alpha_t : \hat{\mathcal{Z}} \times \mathcal{Q} \to [0,1]$. The minimum is over all $v \in \mathcal{Q}$.

Assumption 1:

(i) $\alpha_t(\hat{z}, Q) = 0$ unless $(\hat{z}_t^N, Q_t) = (\hat{z}, Q)$. Also,

$$\alpha_t(\hat{z}_t^N, Q_t) = \frac{1}{1 + \sum_{k=0}^t \mathbf{1}_{\{\hat{z}_k^N = \hat{z}_t^N, Q_k = Q_t\}}}.$$

(ii) The policy γ' used to collect the realizations chooses the quantizers $(Q_t)_{t\geq N}$ independently and uniformly from $\mathcal{Q}$.

(iii) The fixed prior $\hat{\pi}$ used in the finite-memory construction is the unique invariant distribution of the source $(x_t)_{t\geq 0}$.

Theorem 4: 1) Under Assumption 1, $(Q_t)_{t\geq N}$ converges almost surely to a limit Q^* satisfying

$$Q^*(\hat{z}, Q) = c(\hat{z}, Q) + \beta \int_{\hat{\mathcal{Z}}} Q(\hat{z}_1^N, Q)\hat{\eta}(d\hat{z}_1^N|\hat{z}, Q)$$

Furthermore, define $\hat{\gamma}^*$ as $\hat{\gamma}^*(\hat{z}) = \operatorname{argmin}_v Q^*(\hat{z}, v)$. Then $\hat{\gamma}^*$ satisfies the optimality equation (5) for $t \geq N$.

2) Denote by $\widetilde{\gamma}^*$ the extension of $\hat{\gamma}^*$ to $\mathcal{Z}$, as in (6). Then $\widetilde{\gamma}^*$ satisfies

$$\mathbf{E}_{\pi_0}^\gamma \left[\left|J_\beta(z_N^N, \widetilde{\gamma}^*) - J_\beta^*(z_N^N)\right|\right] \leq \frac{2||d||_\infty}{1-\beta} \sum_{t=0}^{\infty} \beta^t L_t^N,$$

where γ acts on the first N steps and L_t^N is as in (8).

VIII. Simulation

We now give an example. Further examples are given in the full paper [25].

We take $\beta = 0.99$, the distortion function $d(x, \hat{x}) = (x - \hat{x})^2$, and use a uniform measure for $\hat{\pi}$. The discounted cost is approximated by running a simulation over $t = 0$ to $t = 10^5$. Consider a source with transition kernel and channel $O(q'|q)$, with $\mathcal{M} = \mathcal{M}' = \mathbb{X}$, given by

$$T = \begin{pmatrix} \frac{1}{2} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}, \qquad O = \begin{pmatrix} \frac{7}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} \\ \frac{1}{10} & \frac{7}{10} & \frac{1}{10} & \frac{1}{10} \\ \frac{1}{10} & \frac{1}{10} & \frac{7}{10} & \frac{1}{10} \\ \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{7}{10} \end{pmatrix}$$

We have that $\delta(T) = \frac{2}{3}$ and $\delta(O) = \frac{4}{10}$, so we can apply Corollary 2 with $\alpha = (1 - \frac{2}{3})(2 - \frac{4}{10}) = \frac{8}{15}$. In such a setup (where $\mathbb{X} = \mathcal{M}$ and the channel is symmetric), it was shown in [6] that "memoryless" encoding (i.e. where $q_t = x_t$) is optimal. We compare our algorithm against this optimal policy, shown in Figure 1.

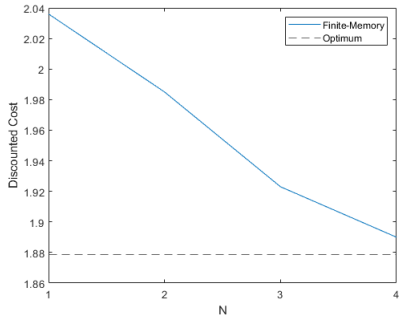

Fig. 1. The optimum is approached as N increases.

Concluding Remarks

As noted in Section VI, Dobrushin-type conditions prioritize uninformative kernels; and more relaxed stability conditions are possible [22]. Furthermore, our analysis is also naturally applicable to continuous alphabet sources. We intend to generalize our results in these directions.

References

[1] C. Shannon, "The zero error capacity of a noisy channel," *IRE Transactions on Information Theory*, vol. 2, no. 3, pp. 8–19, 1956.

[2] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. Cambridge University Press, 2011.

[3] J. M. Ooi, *Coding for Channels with Feedback*. Springer New York, 1998.

[4] D. B. Kurka and D. Gündüz, "DeepJSCC-f: Deep joint source-channel coding of images with feedback," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 178–193, 2020.

[5] J. Xu, B. Ai, N. Wang, and W. Chen, "Deep joint source-channel coding for CSI feedback: An end-to-end approach," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 260–273, 2023.

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

[7] 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.

[8] 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.

[9] T. Javidi and A. Goldsmith, "Dynamic joint source-channel coding with feedback," in *2013 IEEE International Symposium on Information Theory*, 2013, pp. 16–20.

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

[11] V. S. Borkar, S. K. Mitter, and S. Tatikonda, "Optimal sequential vector quantization of Markov sources," *SIAM J. Control and Optimization*, vol. 40, pp. 135–148, 2001.

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

[13] N. Saldi, S. Yüksel, and T. Linder, "On the asymptotic optimality of finite approximations to Markov decision processes with Borel spaces," *Mathematics of Operations Research*, vol. 42, no. 4, pp. 945–978, 2017.

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

[15] A. 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, pp. 1–34, 2023.

[16] L. Cregg, F. Alajaji, and S. Yüksel, "Reinforcement learning for zero-delay coding over a noisy channel with feedback," in *Proceedings of the IEEE Conference on Decision and Control*, 2023, pp. 3939–3944.

[17] A. 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.

[18] L. Cregg, T. Linder, and S. Yüksel, "Reinforcement learning for the near-optimal design of zero-delay codes for Markov sources," *arXiv preprint arXiv:2311.12609*, 2023.

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

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

[21] ——, "Stochastic observability and filter stability under several criteria," *IEEE Transactions on Automatic Control*, pp. 1–16, 2023.

[22] ——, "Robustness to incorrect priors and controlled filter stability in partially observed stochastic control," *SIAM Journal on Control and Optimization*, vol. 60, no. 2, pp. 842–870, 2022.

[23] H. Yu and D. P. Bertsekas, "On near optimality of the set of finite-state controllers for average cost POMDP," *Mathematics of Operations Research*, vol. 33, no. 1, pp. 1–11, 2008.

[24] A. Kara and S. Yüksel, "Near optimality of finite memory feedback policies in partially observed Markov decision processes," *Journal of Machine Learning Research*, vol. 23, no. 11, pp. 1–46, 2022.

[25] L. Cregg, F. Alajaji, and S. Yüksel, "Reinforcement learning for optimal zero-delay coding of Markov sources over noisy channels: Belief quantization vs. finite memory codes," *arXiv*, 2024.

[26] S. Yüksel. (2023) Optimization and control of stochastic systems. [Online]. Available: https://mast.queensu.ca/~yuksel/LectureNotesOnStochasticOptControl.pdf

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

[28] 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.

[29] O. Hernández-Lerma and J. Lasserre, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Springer, 1996.

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

[31] A. Kara and S. Yüksel, "Q-learning for stochastic control under general information structures and non-Markovian environments," *arXiv preprint, arXiv:2311.00123*, 2023.