Reinforcement Learning for Near-Optimal Design of Zero-Delay Codes for Markov Sources

Liam Cregg, *Graduate Student Member, IEEE*, Tamás Linder, *Fellow, IEEE*, and Serdar Yüksel, *Senior Member, IEEE*

***Abstract*— In the classical lossy source coding problem, one encodes long blocks of source symbols that enables the distortion to approach the ultimate Shannon limit. Such a block-coding approach introduces large delays, which is undesirable in many delay-sensitive applications. We consider the zero-delay case, where the goal is to encode and decode a finite-alphabet Markov source without any delay. It has been shown that this problem lends itself to stochastic control techniques, which lead to existence, structural, and general structural approximation results. However, these techniques so far have only resulted in computationally prohibitive algorithmic implementations for code design. To address this problem, we present a practically implementable reinforcement learning design algorithm and rigorously prove its asymptotic optimality. In particular, we show that a quantized Q-learning algorithm can be used to obtain a near-optimal coding policy for this problem. The proof builds on recent results on quantized Q-learning for weakly Feller controlled Markov chains whose application necessitates the development of supporting technical results on regularity and stability properties, and relating the optimal solutions for discounted and average cost infinite horizon criteria problems. These theoretical results are supported by simulations.**

***Index Terms*— Source coding, network control systems, quantization.**

I. Introduction

A. Zero-Delay Lossy Coding

WE CONSIDER the problem of encoding an information source without delay, sending the encoded source over a discrete noiseless channel, and reconstructing the source, also without delay, at the decoder. Hence, the classical block-coding approach is not allowed. Zero-delay coding schemes have many practical applications in emerging fields such as networked control systems (see [1] and references therein for an extensive review and discussion of applications), real-time mobile audio-video systems (as in streaming systems [2]), and real-time sensor networks [3], among other areas.

Manuscript received 21 November 2023; revised 12 April 2024; accepted 5 June 2024. Date of publication 17 June 2024; date of current version 22 October 2024. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada. The work of Liam Cregg was supported by the Queen's University Department of Mathematics and Statistics Summer Research Award. *(Corresponding author: Tamás Linder.)*

The authors are with the Department of Mathematics and Statistics, Queen's University, Kingston, ON K7L 3N6, Canada (e-mail: liam.cregg@queensu.ca; tamas.linder@queensu.ca; yuksel@queensu.ca).

Communicated by Y. Kochman, Associate Editor for Signal Processing and Source Coding.

Digital Object Identifier 10.1109/TIT.2024.3416063

1) Notation:

We will use the subscript t to denote a time-dependent object (as the samples X_t of an information source); any other subscript denotes some other dependency. In particular, the subscript n will be used to denote an underlying quantization parameter, and should not be confused with time. 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.

Random variables will in general be denoted by uppercase letters. We make a few exceptions in order to conform with the corresponding literature on zero-delay coding; in particular, the (random) sequence of quantizers and channel symbols will be denoted by Q_t and q_t respectively. However, they can be distinguished from their realizations by the time subscript. For example, we write $P(Q_t = Q)$ and $P(q_t = q)$. When discussing (time-homogeneous) Markov processes, we will often use the shorthand $P(x'|x) = P(X_{t+1} = x'|X_t = x)$. In the case of multiple Markov chains, which transition probability we mean will be clear from the variable names. For example, if we have two Markov chains $\{X_t\}_{t\geq 0}$ and $\{Y_t\}_{t\geq 0}$, then $P(x'|x)$ and $P(y'|y)$ are the respective transition probabilities.

We use superscripts to denote the product sets, (e.g. $\mathbb{X}^t$ denotes the t-fold product of the source alphabet $\mathbb{X}$), and for random vectors taking values in these sets we use the notation $X_{[0,t-1]} := (X_0, \ldots, X_{t-1})$. Also, we use $\mathcal{P}(\mathbb{X})$ to denote the space of probability measures over the set $\mathbb{X}$ (and we endow this space with the topology of weak convergence).

The source $\{X_t\}_{t\geq 0}$ is a time-homogeneous, discrete-time Markov process taking values in a finite set $\mathbb{X}$ and has transition matrix $P(x'|x)$. We assume that the source is irreducible and aperiodic, and thus admits a unique invariant measure, which we will denote by ζ. We also assume that the distribution of X_0, which we denote by π_0 (this can be different from ζ), is available at the encoder and decoder.

At time $t \geq 0$, the encoded (compressed) information, denoted by q_t, is sent over a discrete noiseless channel with common input and output alphabet $\mathcal{M} := \{1, \ldots, M\}$. The encoder is defined by an encoder policy $\gamma^e = \{\gamma_t^e\}_{t\geq 0}$, where $\gamma_t^e : \mathcal{M}^t \times \mathbb{X}^{t+1} \to \mathcal{M}$, and $q_t = \gamma_t^e(q_{[0,t-1]}, X_{[0,t]})$. Note that

given $q_{[0,t-1]}$ and $X_{[0,t-1]}$, the map $\gamma_t^e(q_{[0,t-1]}, X_{[0,t-1]}, \cdot)$ is a *quantizer* (i.e. a map from $\mathbb{X}$ to $\mathcal{M}$), which we denote by Q_t. We will denote the set of all quantizers from $\mathbb{X}$ to $\mathcal{M}$ by $\mathcal{Q}$. Thus we can view an encoder policy at time t as selecting a quantizer $Q_t \in \mathcal{Q}$ and then encoding (quantizing) X_t as $q_t = Q_t(X_t)$. We call such encoder policies admissible, and denote the set of all admissible encoder policies (sometimes called quantization policies) by Γ^e. Upon receiving q_t, the decoder generates the reconstruction $\hat{X}_t$ without delay, using decoder policy $\gamma^d = \{\gamma_t^d\}_{t\geq 0}$, where $\gamma_t^d : \mathcal{M}^{t+1} \to \hat{\mathbb{X}}$ and where $\hat{\mathbb{X}}$ is a finite reproduction alphabet. Thus we have $\hat{X}_t = \gamma_t^d(q_{[0,t]})$. The set of these admissible decoder policies is denoted by Γ^d.

In general for the zero-delay coding problem, the goal is to minimize the average distortion (cost), given by

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

where $d : \mathbb{X} \times \hat{\mathbb{X}} \to [0, \infty)$ is a given distortion measure and $\mathbf{E}_{\pi_0}^{\gamma^e,\gamma^d}$ is the expectation with initial distribution $X_0 \sim \pi_0$ under encoder policy γ^e and decoder policy γ^d.

It is straightforward to show that, for a fixed encoder policy γ^e, the optimal decoder policy for all $t \geq 0$ is given by

$$\gamma_t^{d*}(q_{[0,t]}) = \operatorname*{argmin}_{\hat{x}\in\hat{\mathbb{X}}} \mathbf{E}_{\pi_0}^{\gamma^e}\left[d(X_t, \hat{x})|q_{[0,t]}\right]. \quad (2)$$

Thus we identify a coding policy with the corresponding encoder policy by assuming that an optimal decoding policy is used for any given encoding policy and will denote $\gamma := \gamma^e$ and $\Gamma := \Gamma^e$. We can then restrict our search to finding optimal encoding policies. With an abuse of notation, we denote

$$J(\pi_0, \gamma) := \inf_{\gamma^d\in\Gamma^d} J(\pi_0, \gamma^e, \gamma^d).$$

The objective is then to minimize $J(\pi_0, \gamma)$ over all Γ. We will denote the optimal cost by

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

We will also consider the discounted cost problem, which is the minimization of

$$J_\beta(\pi_0, \gamma^e, \gamma^d) := \lim_{T\to\infty} \mathbf{E}_{\pi_0}^{\gamma^e,\gamma^d}\left[\sum_{t=0}^{T-1} \beta^t d(X_t, \hat{X}_t)\right]. \quad (3)$$

for some $\beta \in (0, 1)$. As above, we assume an optimal decoder policy and minimize only over the encoder policies, yielding $J_\beta(\pi_0, \gamma)$ and $J_\beta^*(\pi_0) := \inf_{\gamma\in\Gamma} J(\pi_0, \gamma)$. We note that, as opposed to the optimal average distortion $J^*(\pi_0)$, the quantity $J_\beta^*(\pi_0)$ has little importance from a source coding point of view and we only use it as a tool toward designing codes that are near-optimal in the average distortion sense.

We say that a set of policies $\{\gamma\}$ depending on some parameter set is *near-optimal* if for any $\epsilon > 0$, there is some choice of parameters (to be identified explicitly later) such that the resulting policy γ satisfies $J(\pi_0, \gamma) \leq J^*(\pi_0) + \epsilon$.

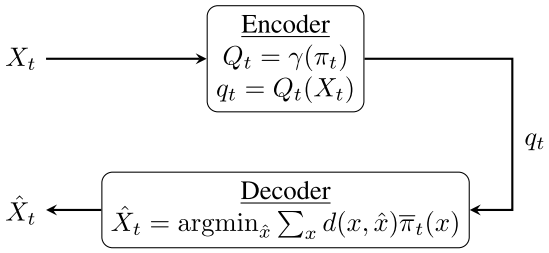

Fig. 1. Block diagram for our zero-delay coding system: X_t is the source sample, q_t is the encoded symbol transmitted through the noiseless channel, and $\hat{X}_t$ is the reconstructed source sample.

B. Literature Review

A number of important results have been obtained in the literature concerning the structure of optimal zero-delay codes, starting with the foundational papers by Witsenhausen [4] and Walrand and Varaiya [5]. In particular, for the finite horizon problem, [4] showed that any encoder policy can be replaced, without performance loss, by one using only $q_{[0,t-1]}$ and X_t to generate q_t. Furthermore, [5] proved a similar result for an encoder policy using only the conditional probability $P(X_t \in \cdot|q_{[0,t-1]})$ and X_t to generate q_t. These results have been further generalized, see e.g., [6], [7], [8], [9], and [10]. In particular, [10] showed the existence of optimal policies in the infinite-horizon case, which we now review.

Recall that $\mathcal{P}(\mathbb{X})$ denotes the space of probability measures over $\mathbb{X}$, and define the conditional probability $\pi_t \in \mathcal{P}(\mathbb{X})$ as the "belief" on X_t given $q_{[0,t-1]}$, i.e.

$$\pi_t(A) := P_{\pi_0}^{\gamma}(X_t \in A|q_{[0,t-1]}) \quad (4)$$

for all measurable $A \subset \mathbb{X}$. Define $\overline{\pi}_t$ similarly, but conditioned on $q_{[0.t]}$, i.e.

$$\overline{\pi}_t(A) := P_{\pi_0}^{\gamma}(X_t \in A|q_{[0,t]}). \quad (5)$$

Note that π_t and $\overline{\pi}_t$ are often called the *predictor* and the *filter*, respectively.

Definition 1: We say an encoder policy $\gamma = \{\gamma_t\}_{t\geq 0}$ is of the *Walrand-Varaiya type* if, at time t, the policy uses only π_t and X_t to generate q_t. That is, γ selects a quantizer $Q_t = \gamma_t(\pi_t)$ and q_t is generated as $q_t = Q_t(X_t)$. Such a policy is called *stationary* if it does not depend on t. We denote the set of stationary Walrand-Varaiya type policies by Γ_{WS}.

Note that π_t and $\overline{\pi}_t$ can be obtained recursively from π_{t-1}, q_{t-1}, and Q_{t-1} (see the update equation (6), see also [51], [52]). Since the initial distribution π_0 is known to both the encoder and decoder, the decoder can track π_t, $\overline{\pi}_t$ and Q_t for all $t \geq 0$. The overall coding scheme for a stationary Walrand-Varaiya encoder is summarized in Figure 1. Recall that the optimal decoder is given by (2), and note that the expectation in (2) is simply an expectation with respect to $\overline{\pi}_t$.

An important fact regarding the Walrand-Varaiya type formulation of zero-delay coding of Markov sources is that the belief process $\{\pi_t\}_{t\geq 0}$ can be considered as a $\mathcal{P}(\mathbb{X})$-valued Markov Decision Process (MDP) with $\mathcal{Q}$-valued control $\{Q_t\}_{t\geq 0}$, see, e.g., [8] or [10] for a proof. This observation was fundamental in deriving the following results.

Proposition 1 [10, Theorem 3]: There exists an optimal policy $\gamma^* \in \Gamma_{\mathrm{WS}}$ for the average cost problem (1). That is, there exists $\gamma^* \in \Gamma_{\mathrm{WS}}$ such that

$$J(\pi_0, \gamma^*) = J^*(\pi_0) \quad \text{for all } \pi_0.$$

Proposition 2 [10, Proposition 2]: For any $\beta \in (0,1)$, there exists an optimal policy $\gamma^*_\beta \in \Gamma_{\mathrm{WS}}$ for the discounted cost problem (3). That is, there exists $\gamma^*_\beta \in \Gamma_{\mathrm{WS}}$ such that

$$J_\beta(\pi_0, \gamma^*_\beta) = J^*_\beta(\pi_0) \quad \text{for all } \pi_0.$$

Despite the key structural results reviewed above, finding an optimal policy for either finite or infinite horizons is difficult. For some very special cases, the optimal solution is known. For example, [5] showed memoryless encoding is optimal when $\mathbb{X} = \mathcal{M}$ and the channel is noisy and symmetric. However, for a general source and channel (or for a noiseless channel, as in this paper), finding an optimal encoding policy is an open problem.

Stochastic control[1] based approaches play an important role in the above structural results. Under a stochastic control framework, [4], [5] used dynamic programming, [10], [14], [15] made use of the value iteration algorithm and the vanishing discount method, whereas [9] used the convex analytic method to obtain structural results for optimal codes. A key component in these results, and one that we will use in showing the near-optimality of our algorithm, is that the zero-delay coding problem can be restated as a Markov Decision Process (MDP) with $\{\pi_t\}_{t\geq 0}$ as the state process. However, there are limitations when using these methods to obtain an optimal solution. In particular, dynamic programming relies on backwards induction from a finite time horizon, which is not applicable for the infinite horizon case. Furthermore, the implementation of the value iteration algorithm requires the computation of certain value functions and conditional expectations, which is practically very challenging due to the probability measure-valued state dynamics.

These challenges, which will be made more explicit in Section III-A, motivate the use of a reinforcement learning approach that we present in this paper. A popular reinforcement learning algorithm, Q-learning [13], [16], [17], [18], [19], [20] is primarily used for fully observed finite space MDPs. This algorithm does not require the knowledge of the transition kernel, or even the cost (or reward) function, for its implementation. In this algorithm, the incurred per-stage cost variable is observed through the simulation of a single sample path. When the state and action spaces are finite, under mild conditions that require that all state-action pairs are visited infinitely often, this algorithm is known to converge to the optimal cost. Recently, this algorithm has been generalized to be applicable for continuous space MDPs (see [21] and the references therein).

In the broader literature related to zero-delay coding, often information theoretic relaxation techniques are used to convexify the non-convex zero-delay optimal quantization problem, which lead to lower bounds on optimal performance, as well as to upper bounds. These include replacing the number of bins constraint with a mutual information constraint, applying the Shannon lower bounding technique, or entropy coding (see, e.g., [22], [23] [24]). Using ergodicity and invariance properties, [25] has constructed time-invariant coding schemes using dithering. A further line of work for linear systems follow the sequential rate-distortion theoretic approach [24], [26], [27], [28], [29], [30]. For coding of Gaussian sources over additive Gaussian channels, some of these results become operational for zero-delay coding [26]; see also [28], [31], and [32].

We note that applying learning theoretic methods in the theory of optimal (lossy) source coding has prior history. A well established line of study of this problem focuses on empirical learning methods for data compression [33], [34], [35], though often limited to i.i.d source models. Our approach here is complementary, since we consider a highly structured coding problem instead of (unstructured) vector quantization and we consider sources with memory. Our analysis leads to near-optimal solutions directly (without learning the source distribution). We also refer to recent research activity in machine learning methods in communications theory (see e.g. [36]); however, our analysis seems to be the first contribution to source coding in the context of reinforcement learning.

Recently, the approach we introduce and analyze in this paper has been extended to noisy channels with feedback in [37]. In [38] an alternative sliding finite window code has been introduced. Compared with the sliding finite window approach of that work, the approach in this paper has the following advantages: (i) quantizing probability measures allows for more relaxed filter stability conditions (e.g., with no geometric filter stability conditions); (ii) it avoids the use of a transient period until the initial memory data is collected. Thus, our approach is complementary and is applicable for a wider class of models. This comes at the cost of higher computational load due to the Bayesian updates at the encoder during the implementation prior to the quantization of probability measures.

Finally, stochastic control and reinforcement learning techniques have also been applied to the "dual" problem of channel coding. For example, [14], [39], and [40] all use an MDP formulation along with dynamic programming to obtain results on channel capacity and capacity-achieving codes. Learning-theoretic results to this end include [41], [42].

C. Contributions

We formulate the zero-delay coding problem so that it is amenable to a reinforcement learning approach. In particular, the MDP associated with our zero-delay lossy coding problem has an uncountable state space (the set of beliefs) and thus has to be discretized (quantized) to apply Q-learning. After posing the problem as an MDP, we build on recent results from [21] to rigorously justify the convergence of a reinforcement learning algorithm to a near-optimal solution

[1] We refer the reader, e.g., to the texts [11], [12], and [13] for an introduction into the theory of stochastic control.

(depending on the discretization on the state space), first for the discounted cost problem, and then for the average cost problem. In particular, [21] showed that, under mild assumptions, a Q-learning algorithm in which the state is quantized converges to the optimal solution as the maximum diameter of the quantization bins for the state space goes to zero. However, the results of [21] cannot be straightforwardly applied to our zero-delay coding problem and there are several additional ingredients needed for our analysis: (a) The convergence and near-optimality was shown in [21] only for the discounted cost criterion problem. Our focus is the average cost setup; this will be addressed via relating discounted cost optimal coding policies to ones that are near-optimal for the average cost. (b) We also need to prove several technical results that are necessary for applying the algorithm in [21], such as the unique ergodicity under an exploration policy which has not been studied for our setup. Specifically, our main contributions are the following:

- We present a reinforcement learning algorithm for the near-optimal design of stationary zero-delay codes (Algorithm 1). As an auxiliary result, we state the near-optimality of the algorithm for the discounted cost problem when the source starts from the invariant distribution (Theorem 1). Then we show that an optimal policy for the discounted cost problem (for sufficiently large discount parameter) can be used to obtain a near-optimal policy for the average cost problem (Theorem 2). This gives, to our knowledge, the first concrete implementation of a provably near-optimal algorithm for the zero-delay coding problem.
- To show the convergence of our algorithm, we prove additional regularity properties of the MDP formulation of the zero-delay coding problem. In particular, we show that the process $\{\pi_t\}_{t\geq 0}$ is stable under the uniform exploration policy, and then deduce unique invariance under this same policy, which is necessary for the application of [21]. Furthermore, unlike in [21], due to the lack of strong recurrence conditions of the predictor process in our setup, additional analysis of the initialization is necessary, which we also provide.
- Finally, Section VI provides simulations comparing Algorithm 1 with the so called omniscent finite-state scalar quantization (O-FSSQ) design algorithm [43, Chapter 14], a heuristic, but effective technique for designing zero-delay lossy codes, to demonstrate the superior performance of Algorithm 1.

The rest of the paper is organized as follows: In Section II we present our reinforcement learning algorithm and in Theorem 1 and Theorem 2 state the near-optimality of the resulting stationary policies for the discounted and average cost problems, respectively, when the source starts from its invariant distribution. Although proofs of the main results are given in Section V, we first have to introduce necessary concepts and auxiliary results that will be needed in these proofs. In particular, in Section III-A we review how the zero-delay coding problem can be turned into a problem involving an MDP, and in Section III-B we present the quantized Q-learning algorithm in [21] in the context of a general MDP. In Section IV we prove that the zero-delay coding MDP meets the necessary assumptions to apply the quantized Q-learning algorithm, in particular the unique ergodicity of our MDP under the uniform exploration policy. Section V contains the proofs of the main results and Section VI presents simulation results. Conclusions are drawn in Section VII, where future research directions are also discussed. Some technical results relating discounted cost optimal and average cost optimal policies are relegated to Appendix A.

II. Near-Optimal Design of Zero-Delay Codes

In order to make the algorithm self-contained, we first introduce some definitions and update equations. The rationale for these will be formalized during the proof of convergence to near-optimality, but we give a high-level justification in this section. Note that the implementation of our algorithm does not require the technical knowledge of MDPs used in the proofs, so we first present our algorithm in its entirely and then prove its near-optimality in the following sections.

Recall the definition of the belief π_t in (4). Under a Walrand-Varaiya type policy, π_{t+1} can be obtained from π_t, q_t, and Q_t via the update equation [9]

$$\pi_{t+1}(x') = \frac{1}{\pi_t(Q_t^{-1}(q_t))} \sum_{x\in Q_t^{-1}(q_t)} P(x'|x)\pi_t(x), \qquad (6)$$

where $Q_t^{-1}(q_t) = \{x \in \mathbb{X} : Q_t(x) = q_t\}$ and $\pi_t(Q_t^{-1}(q_t)) = \sum_{x\in Q_t^{-1}(q_t)} \pi_t(x)$. Note that the encoder and decoder can both track π_t and thus Q_t, so these quantities are known at time t. This update essentially performs a Bayesian update on π_t to compute $\overline{\pi}_t$, then computes $\pi_{t+1}(x') = \sum_x P(x'|x)\overline{\pi}_t(x)$.

Recall that $\mathcal{Q}$ is the set of all quantizers from $\mathbb{X} \to \mathcal{M}$. We wish to compute the "cost" of using a given quantizer Q when the predictor is π. The natural choice is the expected distortion, given Q and π, using γ^{d*} (recall (2)) as the decoder. This yields the following cost function $c : \mathcal{P}(\mathbb{X}) \times \mathcal{Q} \to \mathbb{R}_+$,

$$c(\pi, Q) := \sum_{i=1}^{M} \min_{\hat{x}\in\hat{\mathbb{X}}} \sum_{x\in Q^{-1}(i)} \pi(x) d(x, \hat{x}). \qquad (7)$$

We approximate $\mathcal{P}(\mathbb{X})$ with the following finite set. Let $m := |\mathbb{X}|$. Given a fixed parameter $n \in \mathbb{Z}_+$, we define

$$\mathcal{P}_n(\mathbb{X}) := \Big\{\hat{\pi} \in \mathcal{P}(\mathbb{X}) : \hat{\pi} = \Big(\frac{k_1}{n}, \ldots, \frac{k_m}{n}\Big), \quad k_i \in \mathbb{Z}_+,\ i = 1, \ldots, m\Big\}. \qquad (8)$$

Given any $\pi \in \mathcal{P}(\mathbb{X})$, let $\hat{\pi}$ denote the nearest neighbour of π (in Euclidean distance) in $\mathcal{P}_n(\mathbb{X})$. We note that $\hat{\pi}$ can be effectively calculated using [44, Algorithm 1], which we include as Algorithm 2 in Appendix B for convenience. This algorithm "quantizes" $\pi \in \mathcal{P}(\mathbb{X})$ to its nearest neighbor $\hat{\pi} \in \mathcal{P}_n(\mathbb{X})$.

Finally, consider a $\mathcal{P}(\mathbb{X}) \times \mathcal{Q}$-valued sequence $\{\pi_t, Q_t\}_{t\geq 0}$ and the resulting $\mathcal{P}_n(\mathbb{X}) \times \mathcal{Q}$-valued sequence $\{\hat{\pi}_t, Q_t\}_{t\geq 0}$,

where $\hat{\pi}_t$ is the nearest neighbor of π_t in $\mathcal{P}_n(\mathbb{X})$. The following updates are based on the Q-learning equations in [21]. We define Q_t and α_t, which are both functions from $\mathcal{P}_n(\mathbb{X}) \times \mathcal{Q}$ to $[0, \infty)$, by

$$\begin{aligned} \mathsf{Q}_0(\hat{\pi}, Q) &\equiv 0 \\ \mathsf{Q}_{t+1}(\hat{\pi}_t, Q_t) &= (1 - \alpha_t(\hat{\pi}_t, Q_t))\mathsf{Q}_t(\hat{\pi}_t, Q_t) \\ &\quad + \alpha_t(\hat{\pi}_t, Q_t)[c(\pi_t, Q_t) + \beta \min_{v \in \mathcal{Q}} \mathsf{Q}_t(\hat{\pi}_{t+1}, v)] \end{aligned} \tag{9}$$

$$\mathsf{Q}_{t+1}(\hat{\pi}, Q) = \mathsf{Q}_t(\hat{\pi}, Q) \text{ for all } (\hat{\pi}, Q) \neq (\hat{\pi}_t, Q_t)$$

$$\alpha_t(\hat{\pi}, Q) = \frac{1}{1 + \sum_{k=0}^{t} \mathbf{1}((\hat{\pi}_k, Q_k) = (\hat{\pi}, Q))} \tag{10}$$

where $\mathbf{1}((\hat{\pi}_k, Q_k) = (\hat{\pi}, Q))$ is the indicator function of the pair $(\hat{\pi}, Q)$.

With these definitions, we are now ready to introduce our Q-learning algorithm to find a near-optimal encoding policy $\gamma \in \Gamma_{WS}$.

Algorithm 1 Q-learning for Near-Optimal Zero-Delay Quantization

Require: source alphabet $\mathbb{X}$, channel alphabet $\mathcal{M}$, transition kernel $P(x'|x)$, initial distribution π_0, quantization parameter n, discount factor $\beta \in (0, 1)$

1: Sample X_0 according to π_0
2: Quantize π_0 using Algorithm 3 with parameter n to obtain $\hat{\pi}_0$
3: Randomly select quantizer Q_0 uniformly from $\mathcal{Q}$
4: $q_0 = Q_0(X_0)$
5: **for** $t \geq 0$ **do**
6: Compute $c(\pi_t, Q_t)$ using (7)
7: Sample X_{t+1} according to $P(x'|x)$
8: Compute π_{t+1} using (6)
9: Quantize π_{t+1} using Algorithm 3 with parameter n to obtain $\hat{\pi}_{t+1}$
10: Update Q_{t+1} and α_{t+1} using (9) and (10)
11: Randomly select quantizer Q_{t+1} uniformly from $\mathcal{Q}$
12: $q_{t+1} = Q_{t+1}(X_{t+1})$
13: **end for**

In the above algorithm, the initial distribution π_0 can be arbitrary. Note that the stopping criterion in this algorithm can be any measure of the convergence of Q_t. In our implementation, we stop the algorithm when the pointwise difference between Q_{t+1} and Q_t is below some small threshold, i.e. $\max_{\hat{\pi}, Q} |\mathsf{Q}_{t+1}(\hat{\pi}, Q) - \mathsf{Q}_t(\hat{\pi}, Q)| \leq \epsilon$. For results on convergence time for Q-learning algorithms see e.g. [60]. When choosing Q_{t+1} uniformly from $\mathcal{Q}$, one could exhaustively compute the set of all quantizers (this can be done offline) and choose uniformly from them. An alternative (and equivalent) method is to randomly choose $q = Q(x)$ from $\mathcal{M}$, according to the uniform distribution, for each $x \in \mathbb{X}$.

The following two theorems are our main results. Both are proved in Section V, but these proofs rely on concepts and auxiliary results from Sections III and IV.

The first result shows convergence of Algorithm 1 to a near-optimal policy for the discounted cost (distortion) problem if the source starts from the the unique invariant distribution ζ. Note that in this case $\{X_t\}_{t \geq 0}$ is a stationary and ergodic source. Recall that n determines the fineness of the quantization from $\mathcal{P}(\mathbb{X})$ to $\mathcal{P}_n(\mathbb{X})$.

Theorem 1 (Discounted Distortion): For any $n \geq 1$ and $\beta \in (0, 1)$, the sequence $\{\mathsf{Q}_t\}_{t \geq 0}$ converges almost surely to a limit Q^*. For any $\pi \in \mathcal{P}(\mathbb{X})$, let $\hat{\pi}$ denote the nearest neighbor of π in $\mathcal{P}_n(\mathbb{X})$ and define the encoding policy $\gamma_{\beta,n}(\pi)$ by setting

$$\gamma_{\beta,n}(\pi) = \operatorname*{argmin}_{Q \in \mathcal{Q}} \mathsf{Q}^*(\hat{\pi}, Q). \tag{11}$$

Then, for any $\epsilon > 0$ and $\beta \in (0, 1)$, there exists $N \geq 1$ such that

$$J_\beta(\zeta, \gamma_{\beta,n}) \leq J^*_\beta(\zeta) + \epsilon$$

for all $n \geq N$.

Remarks:

(a) We note that $\mathsf{Q}^*(\pi, Q)$ is the expected value of the discounted cost obtained by a policy that takes "action" Q (i.e., uses the quantizer Q) at an initial state π, and then follows the optimal policy for the rest of the time. Due to the step where the minimum of $\mathsf{Q}^*(\hat{\pi}, Q)$ is considered instead of that of $\mathsf{Q}^*(\pi, Q)$, the encoding policy $\gamma_{\beta,n}(\pi)$ is a piecewise constant function of the actual belief $\pi \in \mathcal{P}(\mathbb{X})$.

(b) In the theorem we have used the limiting Q-value Q^* to derive the desired policy $\gamma_{\beta,n}$. Similar (probabilistic and in-expectation) bounds can be given if we use Q_t for finite (large) t, at the expense of a more involved analysis. See [49] and [50] for details on finite-time analysis for Q-learning.

In the preceding theorem the discounted cost (distortion) is considered, a quantity which has limited significance in source coding. The next theorem, which is the main result of this paper, shows that the policy obtained in Theorem 1, for β close enough to 1 and n large enough is also a near-optimal policy for the average cost (distortion) problem. As in Theorem 1, we assume that $\{X_t\}_{t \geq 0}$ starts from the unique invariant distribution ζ.

Theorem 2 (Average Distortion): Let $\epsilon > 0$. Then for all sufficiently large $\beta \in (0, 1)$, there exists $N(\beta) \geq 1$ such that

$$J(\zeta, \gamma_{\beta,n}) \leq J^*(\zeta) + \epsilon$$

for all $n \geq N(\beta)$, where $\gamma_{\beta,n}$ is the policy obtained in (11) of Theorem 1.

III. Quantized Q-Learning

In this section, we first introduce an MDP formulation of the zero-delay coding problem, which has been studied in depth in [9] and [10], and describe our motivation for a reinforcement learning approach to solve this MDP. Then we review recent results on a quantized Q-learning algorithm in the context of a *general* MDP, including the necessary assumptions. Sections IV and V are then dedicated to proving that these assumptions hold for the zero-delay coding MDP,

allowing us to apply the quantized Q-learning algorithm. The resulting algorithm (quantized Q-learning applied to the zero-delay coding MDP) is then exactly our Algorithm 1.

A. Zero-Delay Coding as a Markov Decision Process (MDP)

It has been shown in [9] and [10] that solving the zero-delay coding problem is equivalent to solving the MDP with state space $\mathcal{P}(\mathbb{X})$, action space $\mathcal{Q}$, transition kernel $P(d\pi'|\pi, Q)$ (which is induced by the update equation (6)), and cost function $c(\pi, Q)$ given in (7). The following is a key property of this MDP.

Definition 2: A transition kernel $P(dz'|z, u)$ is called *weakly continuous* if

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

is continuous in (z, u) for all continuous and bounded f. If an MDP has a weakly continuous transition kernel, we call the MDP *weak Feller*.

Lemma 1 [9, Lemma 11]: For any $\gamma \in \Gamma_{\text{WS}}$, the transition kernel $P(d\pi_{t+1}|\pi_t, Q_t)$ is weakly continuous.

The MDP formulation of the zero-delay coding problem has many analytical advantages. For example, it allows the use of dynamic programming and value iteration methods to prove existence results. However, this representation entails several limitations. In particular, even though our original source $\{X_t\}_{t\geq 0}$ takes finitely many values, admits a unique invariant measure, and has explicit transition probabilities given by $P(x'|x)$, the MDP representation has $\{\pi_t\}_{t\geq 0}$ as its state process, which takes values in the uncountable set $\mathcal{P}(\mathbb{X})$, and it has an analytically complicated transition probability $P(d\pi'|\pi, Q)$ induced by the update equation (6). This makes actual implementation of dynamic programming and value iteration for the computation of optimal policies difficult.

In particular, a traditional approach to obtain an optimal policy for the discounted cost problem is to use the value iteration algorithm given by

$$J_t(\pi) = \min_{Q\in\mathcal{Q}} \Big[c(\pi, Q) + \beta \int_{\mathcal{P}(\mathbb{X})} J_{t-1}(\pi')P(d\pi'|\pi, Q)\Big],$$

for $t \in \{1, 2, \ldots\}$, with $J_0(\pi) = 0$. It can be shown that our MDP satisfies the conditions for the convergence $J_t(\pi_0) \to J^*_\beta(\pi_0)$ as $t \to \infty$ to hold, and the actions obtaining the above minimum converge to an optimal policy [10].

It turns out however that actually computing this value function is difficult; the values clearly cannot be computed directly for each state as the state space is uncountable. An approach would be to quantize the MDP via an approximate model whose solution is near-optimal for the original model (e.g. via [46, Theorem 4.27]). For zero-delay quantization, such an approximate model would require numerical simulations for the computation of transition probabilities, as one would need to place a probability measure on sets of probability measures. Thus, computing the above values is very difficult except in trivial cases. This motivates the use of a reinforcement learning algorithm in which the transition probabilities are not computed or estimated explicitly.

B. Quantized Q-Learning

A common reinforcement learning algorithm for finding optimal policies is Q-learning, in which empirical value functions are recursively updated based on observed realizations of the state, action, and cost. Such an algorithm is guaranteed to converge to an optimal policy for the discounted cost problem, but only in situations where the state and action spaces are finite (among other mild assumptions, see [16]), and thus it is not applicable in our case.

A solution is "quantized" Q-learning, where we approximate the original MDP using an MDP with a finite state space, and run Q-learning on this model. Recent work [21] and [47] give conditions under which the resulting policy is near-optimal for the original MDP. We note that such a quantization strategy is not limited to a Q-learning approach. For example, [48] considers a quantized value iteration approach, but while this solves the issue of the uncountable state space, one must still contend with the difficult state dynamics given by $P(\cdot\,|\pi, Q)$. Furthermore, in such a quantization procedure, one must compute probability measures over the transition kernel itself, which makes the problem even more challenging. A Q-learning approach avoids all of this complexity by learning the values empirically.

First, we state the assumptions that allow for the application of quantized Q-learning to a general MDP. In the proof of our main result, we will later show that these assumptions hold for the zero-delay coding MDP. Consider an MDP with Z-valued state $\{Z_t\}_{t\geq 0}$, U-valued control $\{U_t\}_{t\geq 0}$, transition kernel $P(dz'|z, u)$ and cost function $c : \mathsf{Z} \times \mathsf{U} \to [0, \infty)$

Assumption 1: The MDP has the following properties:

(i) The transition kernel $P(dz'|z, u)$ is weakly continuous.
(ii) The cost function c is continuous and bounded.
(iii) The action space U is finite.
(iv) The state space Z is a compact subset of a Euclidean space.

Let $\{B_i\}_{i=1}^N$ be a partition of Z into compact subsets and let $\mathsf{Y} := \{y_1, \ldots, y_N\}$, where $y_i \in B_i$. We define a *quantizer* on Z as a mapping $f : \mathsf{Z} \to \mathsf{Y}$, such that

$$f(z) = y_i \quad \text{if } z \in B_i.$$

Note that when we first introduced quantizers in Section I, we were considering quantizers of the *information source* $\{X_t\}_{t\geq 0}$. Although the idea is the same here, we emphasize that we are now considering quantization of the state space of an MDP.

We also define the maximum radius of the B_i:

$$d_\infty(\mathsf{Y}, \mathsf{Z}) := \max_{i=1,\ldots,N} \max_{z\in B_i} \|z - y_i\|. \tag{12}$$

We now introduce the quantized Q-learning algorithm. Here, we let $y \in \mathsf{Y}$ and $Y_t := f(Z_t)$. Also, we apply a (possibly randomized) policy $\eta := \{\eta_t\}_{t\geq 0}$, where $\eta_t : \mathsf{Y}^{t+1} \times \mathsf{U}^t \to \mathsf{U}$. We refer to this policy as the *exploration policy*. Finally, we define the Q-factors $\{\mathsf{Q}_t\}_{t\geq 0}$ and the learning rate $\{\alpha_t\}_{t\geq 0}$, where both Q_t and α_t are maps from $\mathsf{Y}\times\mathsf{U}$ to $[0, \infty)$. Note that since Y and U are finite, Q_t and α_t are tabular.

Algorithm 2 Quantized Q-learning (General MDPs) [21]

Initialize Z_0
for $t \geq 0$ **do**
  $U_t \sim \eta_t(Y_{[0,t]}, U_{[0,t-1]})$
  Sample Z_{t+1} according to $P(dz'|z,u)$
  if $(Y_t, U_t) = (y,u)$ **then**
    $\mathsf{Q}_{t+1}(y,u) = (1-\alpha_t(Y_t,U_t))\mathsf{Q}_t(Y_t,U_t) + \alpha_t(Y_t,U_t)[c(Z_t,U_t) + \beta \min_{v\in\mathsf{U}} \mathsf{Q}_t(Y_{t+1},v)]$
  else
    $\mathsf{Q}_{t+1}(y,u) = \mathsf{Q}_t(y,u)$
  end if
end for

As shown in [21] and stated in Theorem 3 below, Algorithm 2 converges to the optimum if the following set of assumptions, together with Assumption 1, are satisfied.

Assumption 2: In Algorithm 2, we have

(i) $\alpha_t(y,u) = \frac{1}{1+\sum_{k=0}^{t} \mathbf{1}((Y_k,U_k)=(y,u))}$.
(ii) Under η, each $(y,u) \in \mathsf{Y}\times\mathsf{U}$ is hit infinitely often almost surely.
(iii) Under η, the state process $\{Z_t\}_{t\geq 0}$ admits a unique invariant measure ϕ.

Theorem 3: [21, Corollary 11]: Under Assumptions 1 and 2, Q_{t+1} in Algorithm 2 converges to a limit Q^*. Furthermore, consider the (deterministic and stationary) policy obtained through

$$\eta^*(z) = \operatorname*{argmin}_{v\in\mathsf{U}} \mathsf{Q}^*(f(z),v).$$

Then,

$$\left|J_\beta(z,\eta^*) - J^*_\beta(z)\right| \to 0 \quad \text{as} \quad d_\infty(\mathsf{Y},\mathsf{Z}) \to 0$$

for all $z \in \mathsf{Z}$. That is, the policy obtained by taking the minimizing actions of Q^* and then making it constant over the quantization bins is near-optimal for fine enough quantization.

Returning to our application, we want to apply this algorithm to the MDP defined in Section III-A. Using the notation of this section, we have $\mathsf{Z} = \mathcal{P}(\mathbb{X})$ and $\mathsf{U} = \mathcal{Q}$, with the cost function (7) and transition kernel $P(d\pi'|\pi,Q)$. We let $\mathsf{Y} = \mathcal{P}_n(\mathbb{X})$ (recall (8)) and show that Assumptions 1 and 2 hold for this setup. The most challenging of these will be verifying that Assumption 2 (iii) holds for our MDP; the next section is dedicated to proving this.

IV. Unique Ergodicity Under a Uniform Exploration Policy

To show the desired result, we will need the following supporting results from the literature of partially observed Markov processes (POMPs). In this section, we assume that the quantizers $\{Q_t\}_{t\geq 0}$ are chosen uniformly from $\mathcal{Q}$, as in Algorithm 1, and call this the *uniform policy*. Unless explicitly stated otherwise, we assume that this uniform policy is used, and so we omit γ from the notation for probabilities and expectations.

A. Predictor and Filter Merging

Recall the definition of the *filter* in (5):

$$\overline{\pi}_t(A) := P_{\pi_0}(X_t \in A | q_{[0,t]}).$$

The filter admits a recursion equation similar to (6) (see e.g. [51], [52]). Note that these recursions are dependent on the initialization of π_0, also called the *prior*. We denote the predictor (respectively, filter) process resulting from the prior $\pi_0 = \nu$ as $\{\pi^\nu_t\}_{t\geq 0}$ (respectively, $\{\overline{\pi}^\nu_t\}_{t\geq 0}$). A common problem in filtering theory is that of *filter stability* (see e.g. [51], [52]), which essentially asks when the process $\{\overline{\pi}^\nu_t\}_{t\geq 0}$ is insensitive to the initialization ν. This will be a crucial tool for proving unique ergodicity. We first provide some necessary definitions.

Definition 3: Let $\nu,\mu \in \mathcal{P}(\mathbb{X})$. The total variation distance is

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

where the supremum is taken over all measurable $f : \mathbb{X} \to [-1,1]$.

Definition 4: We say that the predictor (respectively, filter) process is *stable in total variation almost surely* if for any $\mu,\nu,\kappa \in \mathcal{P}(\mathbb{X})$, we have that P_κ almost surely,

$$\lim_{t\to\infty} \|\pi^\mu_t - \pi^\nu_t\|_{TV} = 0.$$

We will use the following lemmas to deduce predictor stability.

Lemma 2: If the filter process is stable in total variation almost surely, then the predictor process is stable in total variation almost surely.

Proof: Consider the source transition kernel $P(x'|x)$. We have that $\pi^\mu_{t+1}(x') = \sum_x P(x'|x)\overline{\pi}^\mu_t(x)$. By a classic result of Dobrushin [57], this implies that $\|\pi^\mu_{t+1} - \pi^\nu_{t+1}\|_{TV} \leq \|\overline{\pi}^\mu_t - \overline{\pi}^\nu_t\|_{TV}$. The result follows. ∎

Lemma 3 [52, Corollary 5.5]: Let $\{A_t\}_{t\geq 0}$ be a discrete-time Markov chain and $\{B_t\}_{t\geq 0}$ be a stochastic process such that the B_t are conditionally independent given $\{A_t\}_{t\geq 0}$. Also assume $P(B_t|A_{[0,\infty)}) = P(B_t|A_t)$, and that $P(B_t|A_t)$ has the form

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

where $g(a,b)$ is a probability density with respect to the σ-finite measure ψ. If g is strictly positive, and $\{A_t\}_{t\geq 0}$ is aperiodic and Harris recurrent (that is, it visits every state infinitely often with probability one [56, Definition 3.1.3]), then the filter $\overline{\pi}_t(A) := P(A_t \in A|B_{[0,t]})$ is stable in total variation almost surely.

Lemma 4: Under the uniform policy, the filter process $\{\overline{\pi}_t\}_{t\geq 0}$ is stable in total variation almost surely.

Proof: We apply Lemma 3 to $\{X_t\}_{t\geq 0}$ and $\{q_t\}_{t\geq 0}$. Note that under the uniform policy, $P(q_t|X_{[0,\infty)}) = P(q_t|X_t)$, and we have

$$\begin{aligned} &P(q_t = q|X_t = x) \\ &= \sum_Q P(q_t = q|X_t = x, Q_t = Q)P(Q_t = Q|X_t = x) \end{aligned}$$

$$= \sum_{Q} \mathbf{1}(Q(x) = q) P(Q_t = Q | X_t = x)$$
$$= \sum_{\{Q: Q(x)=q\}} P(Q_t = Q | X_t = x).$$

where the second equality follows from the fact that $q = Q(x)$ is deterministic. Now, under the uniform policy, the quantizer Q_t is chosen independently and randomly, so $P(Q_t = Q|X_t = x) = P(Q_t = Q)$. Thus we have

$$P(q_t = q|X_t = x) = \sum_{\{Q: Q(x)=q\}} P(Q_t = Q).$$

Since we are considering the set of all possible quantizers, for any $(x, q) \in \mathbb{X} \times \mathcal{M}$, the set $\{Q : Q(x) = q\}$ is nonempty, and under the uniform policy, $P(Q_t = Q) > 0$ for all $Q \in \mathcal{Q}$. Thus $P(q_t = q|X_t = x) > 0$ for all (x, q). This implies the function g in Lemma 3 is positive. Finally, note that $\{X_t\}_{t\geq 0}$ evolves independently of the encoding policy; it is always irreducible and aperiodic (thus, since $\mathbb{X}$ is finite, it is Harris recurrent and aperiodic). The result follows from Lemma 3. ∎

Lemmas 2 and 4 immediately imply the following:

Corollary 1: Under the uniform policy, the predictor process $\{\pi_t\}_{t\geq 0}$ is stable in total variation almost surely.

Theorem 4: Under the uniform policy, $\{\pi_t\}_{t\geq 0}$ admits a unique invariant measure.

Proof: The proof slightly generalizes an argument presented in [53, Corollary 3]. Throughout, we use the notation $\nu(f) := \int f d\nu$. Note that, under any Walrand-Varaiya type policy (and in particular, under the uniform policy), the processes $\{\pi_t\}_{t\geq 0}$ and $\{(X_t, \pi_t)\}_{t\geq 0}$ are Markov. Let T and S be the transition operators associated with $\{\pi_t\}_{t\geq 0}$ and $\{(X_t, \pi_t)\}_{t\geq 0}$, respectively. Recall that ζ is the unique invariant measure of our source $\{X_t\}_{t\geq 0}$.

First note that since the exploration policy is uniform, the induced transition kernel $P(d\pi'|\pi)$ itself it weakly continuous (i.e., $\int f(\pi')P(d\pi'|\pi)$ is continuous in π). Since every Markov process with a weakly continuous kernel on a compact state space admits an invariant measure (see, e.g., [54]), $\{\pi_t\}_{t\geq 0}$ has an invariant measure. Thus, we are left with proving uniqueness. Assume that $m_1, m_2 \in \mathcal{P}(\mathbb{X} \times \mathcal{P}(\mathbb{X}))$ are two invariant measures for $\{(X_t, \pi_t)\}_{t\geq 0}$. Then their projections on $\mathbb{X}$ are invariant for $\{X_t\}_{t\geq 0}$. Then, by unique invariance of ζ we have

$$m_i(dx, d\nu) = P_{m_i}(d\nu|x)\zeta(dx).$$

We now show that $m_1(F) = m_2(F)$ for each F on a set of measure-determining functions, namely those such that $F(x, \nu) = \phi(x)H(\nu(\phi_1), \ldots, \nu(\phi_l))$, where $\phi \in C(\mathbb{X}), \phi_1, \ldots, \phi_l \in C(\mathbb{X})$, H is bounded and Lipschitz continuous with constant L_H, and $l \in \mathbb{Z}_+$ [53].

By invariance we have, for $i = 1, 2$, that

$$m_i(F) = \int_{\mathbb{X}\times\mathcal{P}(\mathbb{X})} \frac{1}{T} \sum_{t=0}^{T-1} S^t F(x, \nu) P_{m_i}(d\nu|x)\zeta(dx).$$

Thus,

$$|m_1(F) - m_2(F)|$$
$$\leq \int_{\mathbb{X}\times\mathcal{P}(\mathbb{X})\times\mathcal{P}(\mathbb{X})} \frac{1}{T} \sum_{t=0}^{T-1} |S^t F(x, \nu_1) - S^t F(x, \nu_2)| \cdot P_{m_1}(d\nu_1|x) P_{m_2}(d\nu_2|x)\zeta(dx)$$
$$\leq L_H \|\phi\| \int_{\mathbb{X}\times\mathcal{P}(\mathbb{X})\times\mathcal{P}(\mathbb{X})} \frac{1}{T} \sum_{t=0}^{T-1} \mathbf{E}[\sum_{i=1}^{l} |\pi_t^{\nu_1}(\phi_i) - \pi_t^{\nu_2}(\phi_i)|] \cdot P_{m_1}(d\nu_1|x) P_{m_2}(d\nu_2|x)\zeta(dx).$$

Since the predictor process is stable in total variation almost surely, and by the dominated convergence theorem, the last integral converges to zero as $n \to \infty$. Then, the joint process $\{(X_t, \pi_t)\}_{t\geq 0}$ admits at most one invariant measure. Next we show that $\{\pi_t\}_{t\geq 0}$ admits at most one invariant measure.

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

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

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

converges weakly to an invariant measure η_j for $\{(X_t, \pi_t)\}_{t\geq 0}$ [56, Theorem 3.3.1].

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

B. Properties of the Unique Invariant Measure and Implications for Learning

Here we identify some properties of the unique invariant measure guaranteed by Theorem 4, which will play a role in the use of Algorithm 2. Throughout, for any measure $\pi \in \mathcal{P}(\mathbb{X})$, we denote its nearest neighbor in $\mathcal{P}_n(\mathbb{X})$ by $\hat{\pi}$ (recall that this nearest neighbor map is performed using Algorithm 3).

Let η be the uniform policy, and let ϕ be the unique invariant measure for $\{\pi_t\}_{t\geq 0}$ under η, as in Theorem 4. Not every element of $\mathcal{P}_n(\mathbb{X})$ (recall (8)) is hit infinitely often under η. As a trivial example, take the case where $\{X_t\}_{t\geq 0}$ is independent and identically distributed (i.i.d.) with distribution given by π. Then we have $\pi_t(x) = \pi(x)$ for all $t \geq 1$, so that only $\hat{\pi} \in \mathcal{P}_n(\mathbb{X})$ is visited infinitely often.

To address this issue, consider the set $\mathcal{B}^\phi := \{B \in \{B_i\}_{i=1}^N : \phi(B_i) > 0\}$ where $\{B_i\}_{i=1}^N$ is the set of bins of $\mathcal{P}(\mathbb{X})$ under the nearest neighbor map. Then consider the set of all $\hat{\pi} \in \mathcal{P}_n^\phi(\mathbb{X})$ whose corresponding bin has positive measure under ϕ, given by

$$\mathcal{P}_n^\phi(\mathbb{X}) := \left\{\hat{\pi} \in \mathcal{P}_n(\mathbb{X}) : \hat{\pi} \in B \text{ for some } B \in \mathcal{B}^\phi\right\}. \quad (13)$$

Lemma 5: Under the uniform policy η, for every $\hat{\pi} \in \mathcal{P}_n^\phi(\mathbb{X})$, we have $\hat{\pi}_t = \hat{\pi}$ infinitely often almost surely.

Proof: By the pathwise ergodic theorem (e.g., [54, Theorem 5.4.1]) there exists some $\mu \in \mathcal{P}(\mathbb{X})$ such that for all measurable and bounded $g : \mathcal{P}(\mathbb{X}) \to \mathbb{R}$,

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

P_μ almost surely as $N \to \infty$. But by Corollary 1, this implies that

$$\frac{1}{N}\sum_{t=0}^{N-1} g(\pi_t^\nu) \to \int g(\pi)\phi(d\pi) \tag{14}$$

P_κ almost surely for all ν, κ. Now let g be the indicator function of the quantization bin of the nearest neighbor map f corresponding to some $\hat{\pi} \in \mathcal{P}_n^\phi(\mathbb{X})$; that is, $g(\pi) = 1$ if $f(\pi) = \hat{\pi}$ and $g(\pi) = 0$ otherwise. Then (14) implies that any bin which has positive measure under ϕ must be hit infinitely often almost surely by π_t^ν. But this is exactly how we defined $\mathcal{P}_n^\phi(\mathbb{X})$; it is the set of all $\hat{\pi}$ whose bins have positive measure under ϕ. Therefore every $\hat{\pi} \in \mathcal{P}_n^\phi(\mathbb{X})$ is hit infinitely often almost surely. Note that ν was arbitrary, so this holds regardless of the initialization $\pi_0 = \nu$. ∎

Now that we have identified a set $\mathcal{P}_n^\phi(\mathbb{X})$ that is hit infinitely often (and thus, one where Assumption 2 (ii) is valid), we claim that we can restrict ourselves to this set without any loss of optimality. This is formalized in the following lemma and corollary.

Lemma 6: Let $\pi_0 \sim \kappa$, where $\kappa \ll \phi$ (κ is absolutely continuous with respect to ϕ) and let $\gamma \in \Gamma_{WS}$. Then for all $\hat{\pi}$ that are reachable from $\hat{\pi}_0$ under γ (that is, such that $P^\gamma(\hat{\pi}_t = \hat{\pi} | \pi_0 \sim \kappa) > 0$ for some t), we have $\hat{\pi} \in \mathcal{P}_n^\phi(\mathbb{X})$.

Proof: By invariance of ϕ under the uniform policy, we have that for any $t \geq 0$,

$$\phi(B') = \frac{1}{t}\sum_{k=0}^{t-1}\sum_{\overline{Q}\in\mathcal{Q}^t} \frac{1}{|\mathcal{Q}^t|}\mathbf{1}(Q_{[0,t-1]} = \overline{Q}) \cdot \int \phi(d\pi) P(\pi_k \in B' | \pi_0 = \pi, Q_{[0,t-1]} = \overline{Q}),$$

where B' is the bin of $\hat{\pi}'$, and the second sum is over all $\overline{Q} \in \mathcal{Q}^t$, i.e. over every possible realization of $Q_{[0,t-1]}$. Now assume that $\kappa(A) = \frac{\phi(A)}{\phi(B)}$ for all $A \subset B$, where B is some bin of the nearest neighbor map with $\phi(B) > 0$; that is, κ is the restriction of ϕ on B. Then given any sequence of quantizers $Q_{[0,t-1]} = \overline{Q}$, we have

$$\begin{aligned} &P(\hat{\pi}_t = \hat{\pi}' | \pi_0 \sim \kappa, Q_{[0,t-1]} = \overline{Q}) \\ &= P(\pi_t \in B' | \hat{\pi}_0 = \hat{\pi}, Q_{[0,t-1]} = \overline{Q}) \\ &= \int \kappa(d\pi) P(\pi_t \in B' | \pi_0 = \pi, Q_{[0,t-1]} = \overline{Q}). \end{aligned}$$

Since the above holds for any sequence of quantizers, for any $\gamma \in \Gamma_{WS}$ we have

$$P^\gamma(\hat{\pi}_t = \hat{\pi}' | \pi_0 \sim \kappa) \ll \phi(B').$$

Furthermore, the above holds for any $\kappa \ll \phi$ since the order of absolute continuity holds. Thus any reachable $\hat{\pi}$ must be such that its bin B satisfies $\phi(B) > 0$, and thus $\hat{\pi} \in \mathcal{P}_n^\phi(\mathbb{X})$. ∎

Corollary 2: Let $\pi_0 = \pi^*$, where $\pi^* \in \text{supp}(\phi)$ and $\text{supp}(\phi)$ denotes the support of ϕ in $\mathcal{P}(\mathbb{X})$, and let $\gamma \in \Gamma_{WS}$. Then for all π that are reachable from π_0 under γ, we have either (i) $\pi \in B$ for some $B \in \mathcal{B}^\phi$, or (ii) π is on the boundary of some $B \in \mathcal{B}^\phi$.

Proof: Let $\{N_m\}_{m\geq 0} \subset \mathcal{P}(\mathbb{X})$ be a sequence of open balls centered at π^* such that $\bigcap_{m=0}^\infty N_m = \pi^*$ and let $\{\kappa_m\}_{m\geq 0} \subset \mathcal{P}(\mathcal{P}(\mathbb{X}))$ be defined as $\kappa_m(A) = \frac{\phi(A)}{\phi(N_m)}$ for all $A \subset N_m$; that is, κ_m is the restriction of ϕ to N_m. Note that by definition, $\phi(N_m) > 0$ and $\kappa_m \ll \phi$ for all m. We also have by weak continuity of $P(d\pi'|\pi, Q)$ that $P(d\pi_t | \pi_0 \sim \kappa_m, Q_{[0,t-1]} = \overline{Q})$ converges weakly to $P(d\pi_t | \pi_0 = \pi^*, Q_{[0,t-1]})$.

Now let $B' \subset \mathcal{P}(\mathbb{X})$ be open. By the Portmanteau theorem (e.g., [55, Theorem 11.1.1]), we have that $\liminf_{m\to\infty} P(\pi_t \in B' | \pi_0 \sim \kappa_m, Q_{[0,t-1]}) \geq P(\pi_t \in B' | \pi_0 = \pi^*, Q_{[0,t-1]})$. By the same argument as the previous lemma, this implies that $P^\gamma(\pi_t \in B' | \pi_0 = \pi^*) \ll \phi(B')$ for any $\gamma \in \Gamma_{WS}$.

Now assume that π is reachable, and take some open ball $N_\epsilon(\pi)$ around π. The above argument implies that $\phi(N_\epsilon(\pi)) > 0$. Since this holds for arbitrary ϵ, it must be that either (i) or (ii) holds. ∎

To summarize, the previous lemma and corollary tell us the following:

1) If our initial measure π_0 is sampled according to a distribution which is absolutely continuous with respect to ϕ, then with probability one the $\{\hat{\pi}_t\}_{t\geq 0}$ process will stay in $\mathcal{P}_n^\phi(\mathbb{X})$.
2) If instead our initial measure π_0 is deterministically set to an element of the support of ϕ, then with probability one the $\{\pi_t\}_{t\geq 0}$ process will stay in $\mathcal{P}_n^\phi(\mathbb{X})$ up to the tie breaking on decision boundaries of $\mathcal{P}_n(\mathbb{X})$.

In practice, the decision boundaries of $\mathcal{P}_n(\mathbb{X})$ are known to us, and furthermore the elements of $\mathcal{P}_n^\phi(\mathbb{X})$ will also be clear after running Algorithm 1 for a sufficiently long time, as these will be the bins that are hit most often. Accordingly, one can easily modify the decision boundaries such that any reachable $\hat{\pi}$ is in $\mathcal{P}_n^\phi(\mathbb{X})$. We assume that such a modification has been done, and henceforth restrict our near-optimal policies to only $\mathcal{P}_n^\phi(\mathbb{X})$. Finally, we show that one such initialization for $\pi_0 \in \text{supp}(\phi)$ is given by ζ, the invariant distribution of the source.

Lemma 7: We have $\zeta \in \text{supp}(\phi)$, where $\text{supp}(\phi)$ denotes the support of ϕ in $\mathcal{P}(\mathbb{X})$.

Proof: Consider some open neighborhood of radius δ containing ζ, say $N_\delta(\zeta) \subset \mathcal{P}(\mathbb{X})$. Now consider a totally "uninformative" quantizer $Q \in \mathcal{Q}$; that is $Q(x) = i$ for all $x \in \mathbb{X}$ and some $i \in \mathcal{M}$. Via the update equation (4), if $Q_t = Q$, we have that $\pi_{t+1} = \pi_t P$, where we have used the matrix notation $P(i,j) := P(X_{t+1} = j | X_t = i)$. By a classical result of Dobrushin [57], there exists some $T > 0$ such that for all $\pi \in \mathcal{P}(\mathbb{X})$, $\pi P^T \in N_\delta(\zeta)$. Thus, if $Q_t =$ for $t = 0, \ldots, T-1$, then we have $\pi_T \in N_\delta(\zeta)$.

But under the uniform policy, we have some fixed positive probability of choosing Q, say $P(Q_t = Q) = p > 0$. In particular, for all $t \geq T$, $P(\pi_t \in N_\delta(\zeta) | \pi_0 = \pi) \geq P(Q_{[t-T,t-1]} = (Q, Q, \ldots, Q)) = p^T$. This implies that

ζ is "accessible" (see [58, Definition 2.1]) and hence that $\zeta \in \text{supp}(\phi)$ [58, Lemma 2.2]. ∎

V. Proofs of Main Results

We are now ready to prove Theorems 1 and 2 with the aid of the auxiliary results developed in Sections III and IV. In particular, we show that Assumptions 1 and 2 hold for our quantized MDP with components

$$\begin{aligned} &\mathsf{Z} = \mathcal{P}(\mathbb{X}),\ \mathsf{Y} = \mathcal{P}_n^{\phi}(\mathbb{X}),\ \mathsf{U} = \mathcal{Q}, \\ &P = P(\cdot\,|\pi, Q),\ c = c(\pi, Q), \end{aligned} \tag{15}$$

from which the proofs will follow. We also let $d_\infty = d_\infty(\mathsf{Y}, \mathsf{Z})$ for notational simplicity, where $d_\infty(\mathsf{Y}, \mathsf{Z})$ was defined in (12).

Let f map $\pi \in \mathcal{P}(\mathbb{X})$ to its nearest neighbor $\hat{\pi} \in \mathcal{P}_n(\mathbb{X}) \subset \mathcal{P}(\mathbb{X})$. Note that we consider the smaller set $\mathcal{P}_n^{\phi}(\mathbb{X})$ rather than all of $\mathcal{P}_n(\mathbb{X})$, but this is without any loss when we start from $\pi_0 = \zeta$ by Corollary 2 and Lemma 7.

Lemma 8: The (quantized) MDP defined by the components in (15) satisfies Assumptions 1 and 2.

Proof: We begin with Assumption 1: (i) holds by Lemma 1. (ii) is clear from our definition of c in (7). Since $\mathbb{X}$ and $\mathcal{M}$ are finite, $\mathcal{Q}$ is finite, so (iii) holds. Finally, $\mathbb{X}$ is finite, so $\mathcal{P}(\mathbb{X})$ is compact, so (iv) holds.

We now show Assumption 2: (i) holds in Algorithm 1 based on the update equation (10). (ii) holds from Lemma 5 and from the uniform selection of $Q_{t+1} \in \mathcal{Q}$. Finally, (iii) holds due to Theorem 4. ∎

We can now prove our main results.

Proof: [Proof of Theorem 1] Algorithm 1 is simply the quantized Q-learning algorithm in Section III applied to the quantized zero-delay coding MDP in (15), and using the uniform policy for η. Note that by definition of $\mathcal{P}_n(\mathbb{X})$, for any $\pi = (p_1, \ldots, p_m) \in \mathcal{P}(\mathbb{X})$ and corresponding $\hat{\pi} = (\hat{p}_1, \ldots, \hat{p}_m) \in \mathcal{P}_n(\mathbb{X})$, we have $|p_i - \hat{p}_i| \le \frac{1}{n}$ for $i = 1, \ldots, n$, so the maximum radius of the quantization bins satisfies

$$\max_{\pi \in \mathcal{P}(\mathbb{X})} \min_{\hat{\pi} \in \mathcal{P}_n(\mathbb{X})} \|\pi - \hat{\pi}\|_2 = O\Big(\frac{1}{n}\Big) \to 0,$$

so we have $d_\infty \to 0$ as $n \to \infty$. Then by Lemma 8 we can apply Theorem 3 when $\pi_0 = \zeta$. Thus for any $\beta \in (0, 1)$ and $\epsilon > 0$ we can choose N such that for all $n \ge N$ in Algorithm 1,

$$J_\beta(\zeta, \gamma_{\beta,n}) \le J_\beta^*(\zeta) + \epsilon \tag{16}$$

as claimed. ∎

Proof: [Proof of Theorem 2] We will use Theorem 5 in Appendix A that, loosely speaking, states that if a policy is near-optimal for a discount factor β close enough to 1, then this policy is also near-optimal for the average cost. Let us verify that the conditions of the theorem are met when $\mathsf{Z} = \mathcal{P}(\mathbb{X})$, $\mathsf{U} = \mathcal{Q}$, $P(\cdot\,|z, u) = P(\cdot\,|\pi, Q)$, and $c(z, u) = c(\pi, Q)$. Indeed, (i)-(iii) are met by the definition of $c(\pi, Q)$ since $\mathcal{P}(\mathbb{X})$ is the standard probability simplex in $\mathbb{R}^{|\mathbb{X}|}$ and $\mathcal{Q}$ is finite. Condition (iv) is true by Lemma 1, and (v) is holds by [10, Lemma 1], which states that for any $\pi, \pi' \in \mathcal{P}(\mathbb{X})$,

$$\big|J_\beta^*(\pi) - J_\beta^*(\pi')\big| \le K\rho_1(\pi, \pi') \le K|\mathbb{X}|,$$

where K is some constant and ρ_1 is the L_1 Wasserstein distance on $\mathcal{P}(\mathbb{X})$ [59].

Now let $\epsilon > 0$ and let $N(\beta) \ge 1$ be such that the policy $\gamma_{\beta,n}$ in Theorem 1 satisfies $J_\beta(\zeta, \gamma_{\beta,n}) \le J_\beta^*(\zeta) + \frac{\epsilon}{2(1-\beta)}$ for all $n \ge N(\beta)$. Then we can apply Theorem 5 with ϵ replaced by $\epsilon/2$ and δ replaced by $\frac{\epsilon}{2(1-\beta)}$ to conclude that for all $\beta \in (0, 1)$ close enough to 1, the average cost performance of $\gamma_{\beta,n}$ satisfies

$$J(\zeta, \gamma_{\beta,n}) \le J^*(\zeta) + \epsilon$$

for all $n \ge N(\beta)$, which completes the proof. ∎

Remark: Note that due to the way the discounted cost scales with β, the policy $\gamma_{\beta,n}$ in the proof does not have to be ϵ-optimal for the discounted cost, but rather it can be chosen to be only $\frac{\epsilon}{2(1-\beta)}$-optimal.

VI. Simulations

In the simulations we use mean-squared error (MSE) $d(x, \hat{x}) = (x - \hat{x})^2$ as our distortion measure and when using Algorithm 1 we fix the discount factor at $\beta = 0.9999$. We let $\pi_0 = \zeta$, so that Theorem 2 is applicable. When testing algorithm performance, we take the average distortion over 10^6 samples, which we denote by D, and calculate the signal-to-noise ratio (SNR) according to

$$\text{SNR} = 10 \log_{10}\left(\frac{\text{var}(X)}{D}\right),$$

where $\text{var}(X)$ is the variance of the stationary Markov source $\{X_t\}_{t \ge 0}$. We plot the SNR for varying quantizer rates, which is given by $R = \log_2(|\mathcal{M}|)$. Finally, in all of the simulations we take $\hat{\mathbb{X}} = \mathbb{X} \subset \mathbb{R}$.

On Algorithm Complexity and Time to Convergence: Note the set $\mathcal{P}_n(\mathbb{X})$, and thus the state space of our quantized MDP, has size $|\mathcal{P}_n(\mathbb{X})| = \binom{n+|\mathbb{X}|-1}{|\mathbb{X}|-1}$ [44]. Furthermore, our action space, if we include every possible quantizer from $\mathbb{X} \to \mathcal{M}$, has size $\mathcal{Q} = |\mathcal{M}|^{|\mathbb{X}|}$. Although these both grow quickly in their respective parameters, we note that the actual utilized state and action spaces tend to be much smaller. Indeed, $\mathcal{P}_n^{\phi}(\mathbb{X})$ tends to be much smaller than $\mathcal{P}_n(\mathbb{X})$; as an extreme case, for an independent and identically distributed (i.i.d.) source, $\mathcal{P}_n^{\phi}(\mathbb{X})$ has only one element. Also, many quantizers do not need to be considered (for example, those with empty bins, and for certain distortion measures those with non-convex bins). Thus, the actual number of states and convergence time will generally be much less than the theoretical upper bound. Note that for linear learning rates (as in our algorithm), the required number of iterations for convergence is polynomial in $|\mathcal{P}_n(\mathbb{X})| \times |\mathcal{Q}|$ (see [60] for related bounds on convergence time for Q-learning).

A. Finite-Alphabet Markov Sources

For finite-state Markov sources, we compare Algorithm 1 against an omniscient finite-state scalar quantizer (O-FSSQ) [43, Chapter 14] as this method seems to be the most competitive among zero-delay code designs for Markov sources. In this algorithm, one fixes a quantizer V from $\mathbb{X}$ to $\{1, \ldots, K\}$, where $\{1, \ldots, K\}$ is a finite

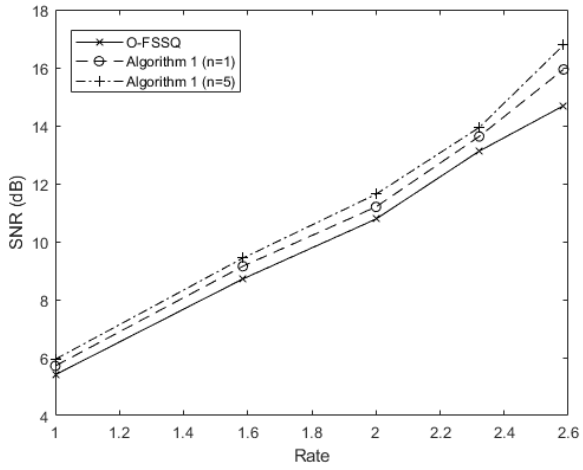

Fig. 2. Comparison of Algorithm 1 with O-FSSQ for finite Markov source.

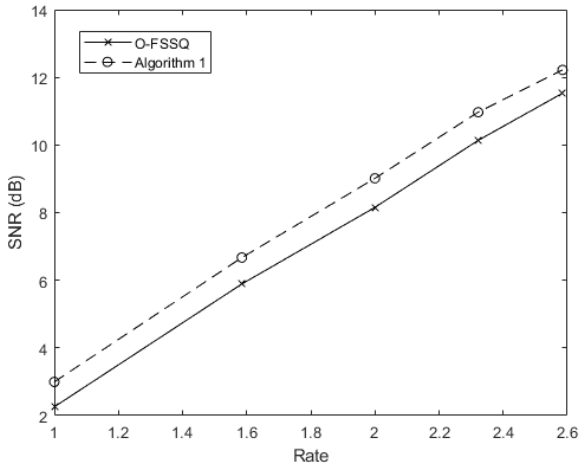

Fig. 3. Comparison of Algorithm 1 with O-FSSQ for Gauss-Markov source.

set of "states". A sequence of training data is sorted into subsequences based on the state of the preceding data point. A Lloyd-Max quantizer [61] is then designed for each of these subsequences. However, in implementation the decoder does not know exactly the state of the preceding data point, so it uses $V(\hat{X}_t)$, rather than $V(X_t)$, as the state. That is, if $V(\hat{X}_t) = i \in \{1, \dots, K\}$, then to quantize X_{t+1} one uses the Lloyd-Max quantizer corresponding to the ith subsequence. The O-FSSQ is known to perform very well for many common Markov sources and on real-world sampled data.

The O-FSSQ employs a strategy similar to Algorithm 1 in the sense that it uses a finite set of "states" at time $t-1$ to select a quantizer at time t. While simple to implement and relatively fast to train, the O-FSSQ design is based on heuristic principles and has no guarantee for optimality or near-optimality, unlike our Algorithm 1. Note that since $\hat{\mathbb{X}}$ is finite, the O-FSSQ is limited to using at most $K = |\hat{\mathbb{X}}|$ states. On the other hand, in Algorithm 1 we can let the number of states be larger in order to obtain a better performance.

In the simulations we consider a stationary Markov source with common source and reproduction alphabet $\mathbb{X} = \hat{\mathbb{X}} = \{1, \dots, 8\}$. The transition matrix and the unique invariant distribution are given below. The transition matrix was randomly (uniformly) chosen from the set of all $|\mathbb{X}| \times |\mathbb{X}|$ transition matrices (in which case the invariant distribution is unique with probability one). In Fig. 2 we plot SNR values for $|\mathcal{M}| = 2, \dots, 6$, so the rates range from $R = 1$ to $R = \log_2(6)$. We include an O-FSSQ designed with $K = 8$ states, using 10^6 training samples. We include two comparisons with Algorithm 1. The first is with $n = 1$, which yields 8 states. The second is with $n = 5$, which gives $\binom{12}{7} = 792$ possible states, but at most 30 are actually utilized in the final design.

1) Transition matrix of source in Figure 1:

$$\begin{pmatrix} 0.1331 & 0.0824 & 0.0311 & 0.2131 & 0.2623 & 0.0714 & 0.0417 & 0.1645 \\ 0.1207 & 0.1501 & 0.1268 & 0.1974 & 0.0952 & 0.0862 & 0.1870 & 0.0362 \\ 0.2320 & 0.0491 & 0.1770 & 0.1476 & 0.1530 & 0.1691 & 0.0215 & 0.05043 \\ 0.0162 & 0.1930 & 0.2511 & 0.1935 & 0.0688 & 0.1280 & 0.0893 & 0.0597 \\ 0.0420 & 0.1496 & 0.1130 & 0.0478 & 0.1073 & 0.2345 & 0.0692 & 0.2363 \\ 0.1382 & 0.1720 & 0.1378 & 0.1369 & 0.0396 & 0.1923 & 0.1383 & 0.0445 \\ 0.1710 & 0.2153 & 0.1579 & 0.0366 & 0.1530 & 0.1144 & 0.0439 & 0.1075 \\ 0.1292 & 0.0534 & 0.1309 & 0.0315 & 0.2837 & 0.2617 & 0.0103 & 0.0988 \end{pmatrix}.$$

2) Invariant Distribution:

$$(0.1211 \;\; 0.1326 \;\; 0.1416 \;\; 0.1328 \;\; 0.1360 \;\; 0.1580 \;\; 0.0806 \;\; 0.0973).$$

Note that even when using the same number of states, Algorithm 1 outperforms the O-FSSQ. At rate $R = 1$, the SNR gain over the O-FSSQ is 0.31 dB for $n = 1$ and 0.53 dB for $n = 5$. At rate $R = \log_2(6) = 2.58$, the SNR gain over the O-FSSQ is 1.25 dB for $n = 1$ and 2.09 dB for $n = 5$. Moreover, Algorithm 1 can be used with a larger number of states, giving performance gains. Of course, this comes at a cost of overall codebook size, as a different quantizer/codebook must be stored for each different state.

B. Continuous Markov Sources

While the mathematical analysis presented in the paper does not cover the continuous case, as future work we intend to develop a rigorous treatment of the continuous source setup. The weak Feller and ergodicity results follow as before, building on the analysis in [9] that used the convex analytic method to obtain structural results for continuous space Markov models; however the unbounded cost will require additional analysis. Nonetheless, in this section, we demonstrate that the algorithm is also suitable for continuous sources as quantization of the probability measures can be done in a variety of efficient ways, facilitating reinforcement learning.

Notably, [62] and [63, Section V.C] propose a scheme in which a probability measure is first approximated by one with finite support, then further quantized using Algorithm 3. Under certain assumptions, this was shown to be an efficient method for quantizing probability measures under a Wasserstein metric, and consequently, the weak convergence topology.

In particular for our algorithm, one computes $\hat{\pi}_t \in \mathcal{P}_n(\mathbb{X})$ by first approximating $\pi_t \in \mathcal{P}(\mathbb{X})$ by a compactly-supported measure, then approximating this by a finitely-supported one, and finally applying Algorithm 3 to this measure. One similarly approximates the space of quantizers $\mathcal{Q}$ by quantizers on some finite set.

In the simulations, we apply this strategy to a Gauss-Markov source with correlation coefficient of 0.9, and we again consider $|\mathcal{M}| = 2, \dots, 6$. For quantization of $\mathcal{P}(\mathbb{X})$ and

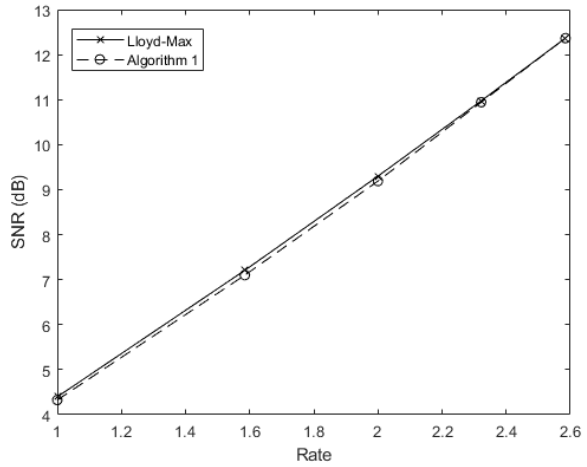

Fig. 4. Comparison of Algorithm 1 with Lloyd-Max for i.i.d. Gaussian source.

$\mathcal{Q}$, we approximate $\mathbb{X} = \mathbb{R}$ by the finite set $\{-6 + 0.05i : i = 0, 1, \ldots, 240\}$ and use $n = 5$ in Algorithm 1. This results in no more than 70 states visited for any given rate, so we compare it to an O-FSSQ using $K = 70$ states, trained on 10^6 samples, where we use a Lloyd-Max quantizer for the state classifier V. Note that in this case the number of states K for the O-FSSQ is not limited, so we only provide a comparison for the same number of states. The results are shown in Fig. 3. At rate $R = 1$, the SNR gain over the O-FSSQ is 0.74 dB and at rate $R = 2.58$, the SNR gain over the O-FSSQ is 0.69 dB.

Note that the question of near-optimality in the continuous case is more intricate, since it depends not only on the parameter n but also on the compact and finite approximations of the support of π_t. That is, even if n is large, Algorithm 1 may perform poorly if the finite approximation is not fine enough.

C. Memoryless Sources

Finally, to demonstrate the effectiveness of the (modified) Algorithm 1, we compare it to the Lloyd-Max quantizer for continuous i.i.d. sources. For such sources, a Lloyd-Max quantizer is only guaranteed to be locally optimal [61]. On the other hand, our Algorithm 1 converges to a globally optimal solution (as $n \to \infty$) so it may outperform a Lloyd-Max quantizer also in the i.i.d. case. However, if the source has a log-concave density, then all local optima coincide with the unique globally optimal quantizer [64], so a Lloyd-Max quantizer designed using the source distribution yields the optimal solution (which is the optimal zero-delay code for the i.i.d. source). To verify near-optimality in this case, we compare Algorithm 1 (with the same quantization parameters as the previous subsection) with a Lloyd-Max quantizer designed using the source distribution, for a standard Gaussian source $X_t \sim N(0, 1)$. The results are shown in Fig. 4. Here the Lloyd-Max design marginally outperforms Algorithm 1 because of the approximation steps used in the latter in the continuous case. The maximum SNR difference across all rates is 0.10 dB, occurring at rate $R = 1.58$. At rate $R = 2.58$, the difference is only 0.002 dB.

VII. Conclusion

We presented a reinforcement learning based algorithm for the design of zero-delay codes for Markov sources. For finite alphabet sources we proved that our Q-learning based algorithm produces zero-delay codes that are optimal as the quantization of the underlying probability space becomes arbitrarily fine, provided the source starts from its invariant distribution. As far as we know, this is the first provably optimal design method for zero-delay coding. The performance of the algorithm was also demonstrated via simulations for finite alphabet as well as continuous sources.

In future work, we aim to rigorously show the near-optimality of our algorithm for continuous alphabet sources with the aid of the quantization scheme in [63, Section V.C], which we already used, in a heuristic way, in our simulations for a continuous source. Furthermore, we are working on generalizing the results to the case where the channel is noisy and the encoder has access to feedback from the channel; most results in our paper seem to go through in this case too, with only a slight modification of the MDP formulation and the update equations.

We are also currently investigating a modified Q-learning approach that uses a finite window of quantizers and quantizer outputs as the state of the code instead of the current belief π_t. The benefits of such an approach include simpler implementation, fewer computations, and an explicit performance bound in terms of the window length. The analysis for such codes becomes quite intricate due to the fact that the notion of predictor stability required for this method is stricter than that we used in this paper, motivating the study of alternative predictor stability conditions.

Appendix A
Supporting MDP Results

In the following, consider an MDP $(\mathsf{Z}, \mathsf{U}, P, c)$, where Z is the state space, U is the action space, $P(dz'|z, u)$ is the transition kernel, and $c : \mathsf{Z} \times \mathsf{U} \to \mathbb{R}_+$ is the cost function.

Assumption 3:
(i) c is continuous, nonnegative, and bounded.
(ii) Z is a compact metric space.
(iii) U is as a compact metric space.
(iv) $P(\cdot\,|z, u)$ is weakly continuous in (z, u).
(v) The family of functions $\{h_\beta : \mathsf{Z} \to \mathbb{R},\ \beta \in (0, 1)\}$, where $h_\beta(z) := J^*_\beta(z) - J^*_\beta(z_0)$ for some fixed $z_0 \in \mathsf{Z}$, is uniformly bounded and equicontinuous.

The following is a standard result in the MDP literature, see e.g. [65, Theorem 3.8], [12, Theorem 5.4.3] for related results.

Lemma 9 [56, Theorems 7.3.3 and 7.3.4]: Consider an MDP which satisfies Assumption 3. Then there exist a constant $g^* \geq 0$ and a measurable function $f : \mathsf{Z} \to \mathsf{U}$ such that the stationary policy $\gamma^* = \{f, f, \ldots\}$ is optimal for the average cost problem and g^* is the optimal value function, i.e.,

$$g^* = J^*(z) = J(z, \gamma^*) \quad \text{for all } z \in \mathsf{Z}.$$

Furthermore, $g^* = \lim_{\beta \uparrow 1}(1 - \beta)J^*_\beta(z)$.

Lemma 10: Let g be a constant and $h : \mathsf{Z} \to \mathbb{R}$, $f : \mathsf{Z} \to \mathsf{U}$ be such that for all $z \in \mathsf{Z}$,

$$g + h(z) \geq c(z, f(z)) + \int h(z')P(dz'|z, f(z)) \tag{17}$$

and

$$\liminf_{t\to\infty} \frac{1}{t}\mathbf{E}_z^\gamma \left[h(Z_t)\right] \geq 0, \tag{18}$$

where $\gamma = \{f, f, \ldots\}$ and $\{Z_t\}_{t\geq 0}$ is the state process under policy γ and arbitrary given initial state $z \in \mathsf{Z}$. Then g is an upper bound to the average cost of policy γ, i.e.,

$$J(z, \gamma) \leq g \quad \text{for all } z \in \mathsf{Z}.$$

Proof: By the law of iterated expectation, we have

$$\mathbf{E}_z^\gamma \left[\sum_{t=1}^{T} h(Z_t)\right] = \mathbf{E}_z^\gamma \left[\sum_{t=1}^{T} \mathbf{E}^\gamma \left[h(Z_t)|Z_{[0,t-1]}, U_{[0,t-1]}\right]\right].$$

By the Markov property and (17),

$$\begin{aligned}
&\mathbf{E}^\gamma \left[h(Z_t)|Z_{[0,t-1]}, U_{[0,t-1]}\right] \\
&\quad = \int h(z')P(dz'|Z_{t-1}, U_{t-1}) \\
&\quad = c(Z_{t-1}, f(Z_{t-1})) + \int h(z')P(dz'|Z_{t-1}, f(Z_{t-1})) \\
&\qquad\qquad - c(Z_{t-1}, f(Z_{t-1})) \\
&\quad \leq g + h(Z_{t-1}) - c(Z_{t-1}, f(Z_{t-1})).
\end{aligned}$$

Substituting this into the previous equation, we obtain

$$\begin{aligned}
&\mathbf{E}_z^\gamma \left[\sum_{t=1}^{T} h(Z_t)\right] \\
&\qquad \leq \mathbf{E}_z^\gamma \left[\sum_{t=1}^{T} g + h(Z_{t-1}) - c(Z_{t-1}, f(Z_{t-1}))\right].
\end{aligned}$$

Rearranging,

$$\mathbf{E}_z^\gamma \left[\sum_{t=1}^{T} c(Z_{t-1}, f(Z_{t-1}))\right] \leq Tg + h(z) - \mathbf{E}_z^\gamma \left[h(Z_T)\right].$$

Dividing by T and taking the limsup,

$$J(z, \gamma) \leq g + \limsup_{T\to\infty} \left(-\frac{1}{T}\mathbf{E}_z^\gamma \left[h(z_T)\right]\right) \leq g,$$

where we use (18) for the last inequality. ∎

The next result shows that if the discount factor β is close enough to 1, a policy that is optimal or near-optimal for the discounted cost, is also near-optimal in the average cost sense.

Theorem 5: Let Assumption 3 hold. Then for every $\epsilon > 0$, there exists a $\beta \in (0, 1)$ such that if a stationary policy $\gamma_\beta = \{f_\beta, f_\beta, \ldots\}$ satisfies for some $\delta \geq 0$,

$$J_\beta(z, \gamma_\beta) \leq J_\beta^*(z) + \delta \tag{19}$$

for all $z \in \mathsf{Z}$, then

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

for all $z \in \mathsf{Z}$, where g^* is the optimal average cost (which is constant by Lemma 9)

Proof: As in Assumption 3, let $h_\beta(z) = J_\beta^*(z) - J_\beta^*(z_0)$ and define $\hat{h}_\beta(z) := J_\beta(z, \gamma_\beta) - J_\beta^*(z_0)$. We will verify that the conditions of Lemma 10 hold with $g = g^* + \epsilon + (1 - \beta)\delta$, $h = \hat{h}_\beta$, and $f = f_\beta$. Indeed, for any $\beta \in (0, 1)$,

$$\begin{aligned}
&\hat{h}_\beta(z) + g^* - (g^* - (1 - \beta)J_\beta^*(z_0)) \\
&\qquad + (1 - \beta) \int \hat{h}_\beta(z')P(dz'|z, f_\beta(z)) \qquad (20) \\
&= \hat{h}_\beta(z) + (1 - \beta)J_\beta^*(z_0) \\
&\qquad + (1 - \beta) \int \hat{h}_\beta(z')P(dz'|z, f_\beta(z)) \\
&= \hat{h}_\beta(z) + (1 - \beta) \int J_\beta(z', \gamma_\beta)P(dz'|z, f_\beta(z)) \\
&= c(z, f_\beta(z)) + \beta \int J_\beta(z', \gamma_\beta)P(dz'|z, f_\beta(z)) - J_\beta^*(z_0) \\
&\qquad + (1 - \beta) \int J_\beta(z', \gamma_\beta)P(dz'|z, f_\beta(z)) \\
&\quad = c(z, f_\beta(z)) + \int \hat{h}_\beta(z')P(dz'|z, f_\beta(z)),
\end{aligned}$$

where the third equality follows from the identity

$$J_\beta(z, \gamma_\beta) = c(z, f_\beta(z)) + \beta \int J_\beta(z', \gamma_\beta)P(dz'|z, f_\beta(z)).$$

Now consider the terms in (20). By Lemma 9, there exists some $\beta_1 \in (0, 1)$ such that

$$\left|g^* - (1 - \beta)J_\beta^*(z_0)\right| \leq \frac{\epsilon}{2} \quad \text{for all } \beta \in [\beta_1, 1). \tag{21}$$

On the other hand, for any $\beta \in (0, 1)$ choose γ_β so that it is δ-optimal, i.e., (19) holds. Then we have

$$\|\hat{h}_\beta\|_\infty \leq \sup_{z\in\mathsf{Z}} \left|J_\beta(z, \gamma_\beta) - J_\beta^*(z)\right| + \|h_\beta\|_\infty \leq \delta + \|h_\beta\|_\infty$$

and therefore

$$\left|(1 - \beta) \int \hat{h}_\beta(z')P(dz'|z, f_\beta(z))\right| \leq (1 - \beta)(\delta + \|h_\beta\|_\infty).$$

By Assumption 3, h_β is uniformly bounded over β, so there exists some $\beta_2 \in (0, 1)$ such that for all $\beta \in [\beta_2, 1)$,

$$\left|(1 - \beta) \int \hat{h}_\beta(z')P(dz'|z, f_\beta(z))\right| \leq (1 - \beta)\delta + \frac{\epsilon}{2}. \tag{22}$$

Thus taking $\beta^* = \max\{\beta_1, \beta_2\}$ in (21) and (22), we have for all $\beta \in [\beta^*, 1)$,

$$\begin{aligned}
&g^* + \epsilon + (1 - \beta)\delta + \hat{h}_\beta(z) \\
&\quad \geq c(z, f_\beta(z)) + \int \hat{h}_\beta(z')P(dz'|z, f_\beta(z)).
\end{aligned}$$

Finally, since $\hat{h}_\beta$ is bounded, (18) is satisfied. Thus by Lemma 10,

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

∎

APPENDIX B
ALGORITHM 3

The following algorithm from [44] is used in our Algorithm 1 to quantize $\pi \in \mathcal{P}(\mathbb{X})$ to its nearest neighbor in $\hat{\pi} \in \mathcal{P}_n(\mathbb{X})$.

Algorithm 3 Predictor Quantization [44, Algorithm 1]

Require: $n \geq 1, \pi = (p_1, \ldots, p_m)$

for $i = 1$**to**m **do**

  $k_i' = \lfloor np_i + \frac{1}{2} \rfloor$

end for

$n' = \sum_i k_i'$

if $n = n'$ **then return** $(\frac{k_1'}{n}, \ldots, \frac{k_m'}{n})$

end if

for $i = 1$**to**m **do**

  $\delta_i = k_i' - np_i$

end for

Sort δ_i s.t. $\delta_{i_1} \leq \ldots \leq \delta_{i_m}$

$\Delta = n' - n$

if $\Delta > 0$ **then**

$$k_{i_j} = \begin{cases} k'_{i_j} & j = 1, \ldots, m - \Delta \\ k'_{i_j} - 1 & j = m - \Delta + 1, \ldots, m \end{cases}$$

else

$$k_{i_j} = \begin{cases} k'_{i_j} + 1 & j = 1, \ldots, |\Delta| \\ k'_{i_j} & j = |\Delta| + 1, \ldots, m \end{cases}$$

end if

return $(\frac{k_1}{n}, \ldots, \frac{k_m}{n})$

REFERENCES

[1] S. Yüksel and T. Başar, *Stochastic Networked Control Systems: Stabilization and Optimization Under Information Constraints*. New York, NY, USA: Birkhäuser, 2013.

[2] S. C. Draper, C. Chang, and A. Sahai, "Lossless coding for distributed streaming sources," *IEEE Trans. Inf. Theory*, vol. 60, no. 3, pp. 1447–1474, Mar. 2014.

[3] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: A survey," *Comput. Netw.*, vol. 38, no. 4, pp. 393–422, 2002.

[4] H. S. Witsenhausen, "On the structure of real-time source coders," *Bell Syst. Tech. J.*, vol. 58, no. 6, pp. 1437–1451, Jul. 1979.

[5] P. Varaiya and J. Walrand, "Causal coding and control for Markov chains," *Syst. Control Lett.*, vol. 3, no. 4, pp. 189–192, Sep. 1983.

[6] D. Teneketzis, "On the structure of optimal real-time encoders and decoders in noisy communication," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 4017–4035, Sep. 2006.

[7] A. Mahajan and D. Teneketzis, "Optimal design of sequential real-time communication systems," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 5317–5338, Nov. 2009.

[8] S. Yüksel, "On optimal causal coding of partially observed Markov sources in single and multi-terminal settings," 2010, *arXiv:1010.4824*.

[9] T. Linder and S. Yuksel, "On optimal zero-delay quantization of vector Markov sources," *IEEE Trans. Inf. Theory*, vol. 60, no. 10, pp. 2975–5991, Dec. 2014.

[10] R. G. Wood, T. Linder, and S. Yüksel, "Optimal zero delay coding of Markov sources: Stationary and finite memory codes," *IEEE Trans. Inf. Theory*, vol. 63, no. 9, pp. 5968–5980, Sep. 2017.

[11] D. P. Bertsekas, *Dynamic Programming and Stochastic Optimal Control*. New York, NY, USA: Academic Press, 1976.

[12] O. Hernândez-Lerma and J. B. Lasserre, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Berlin, Germany: Springer, 1996.

[13] C. Szepesvari, *Algorithms for Reinforcement Learning*. Berlin, Germany: Springer, 2010.

[14] S. Tatikonda and S. Mitter, "The capacity of channels with feedback," *IEEE Trans. Inf. Theory*, vol. 55, no. 1, pp. 323–349, Jan. 2009.

[15] 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 Trans. Inf. Theory*, vol. 68, no. 5, pp. 3474–3488, May 2022.

[16] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, pp. 279–292, May 1992.

[17] J. N. Tsitsiklis, "Asynchronous stochastic approximation and Q-learning," *Mach. Learn.*, vol. 16, no. 3, pp. 185–202, 1994.

[18] W. L. Baker, "Learning via stochastic approximation in function space," Ph.D. dissertation, Harvard Univ., Cambridge, MA, USA, 1997.

[19] C. Szepesvári and M. L. Littman, "A unified analysis of value-function-based reinforcement-learning algorithms," *Neural Comput.*, vol. 11, no. 8, pp. 2017–2060, Nov. 1999.

[20] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*. Belmont, MA, USA: Athena Scientific, 1996.

[21] A. Kara, N. Saldi, and S. Yüksel, "Q-learning for MDPs with general spaces: Convergence and near optimality via quantization under weak continuity," *J. Mach. Learn. Res.*, vol. 24, no. 199, pp. 1–34, 2023.

[22] E. I. Silva, M. S. Derpich, and J. Ostergaard, "A framework for control system design subject to average data-rate constraints," *IEEE Trans. Autom. Control*, vol. 56, no. 8, pp. 1886–1899, Aug. 2011.

[23] E. I. Silva, M. S. Derpich, J. Østergaard, and M. A. Encina, "A characterization of the minimal average data rate that guarantees a given closed-loop performance level," *IEEE Trans. Autom. Control*, vol. 61, no. 8, pp. 2171–2186, Aug. 2016.

[24] P. A. Stavrou, J. Østergaard, and C. D. Charalambous, "Zero-delay rate distortion via filtering for vector-valued Gaussian sources," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 5, pp. 841–856, Oct. 2018.

[25] T. C. Cuvelier, T. Tanaka, and R. W. Heath Jr., "Time-invariant prefix coding for LQG control," 2022, *arXiv:2204.00588*.

[26] R. Bansal and T. Başar, "Simultaneous design of measurement and control strategies for stochastic systems with feedback," *Automatica*, vol. 25, no. 5, pp. 679–694, Sep. 1989.

[27] S. Tatikonda, A. Sahai, and S. Mitter, "Stochastic linear control over a communication channels," *IEEE Trans. Autom. Control*, vol. 49, no. 8, pp. 1549–1561, Oct. 2004.

[28] T. Tanaka, K. K. Kim, P. A. Parrilo, and S. K. Mitter, "Semidefinite programming approach to Gaussian sequential rate-distortion trade-offs," *IEEE Trans. Autom. Control*, vol. 62, no. 4, pp. 1896–1910, Apr. 2017.

[29] M. S. Derpich and J. Ostergaard, "Improved upper bounds to the causal quadratic rate-distortion function for Gaussian stationary sources," *IEEE Trans. Inf. Theory*, vol. 58, no. 5, pp. 3131–3152, May 2012.

[30] P. A. Stavrou and M. Skoglund, "Asymptotic reverse waterfilling algorithm of NRDF for certain classes of vector Gauss–Markov processes," *IEEE Trans. Autom. Control*, vol. 67, no. 6, pp. 3196–3203, Jun. 2022.

[31] P. A. Stavrou, T. Tanaka, and S. Tatikonda, "The time-invariant multidimensional Gaussian sequential rate-distortion problem revisited," *IEEE Trans. Autom. Control*, vol. 65, no. 5, pp. 2245–2249, May 2020.

[32] V. Kostina and B. Hassibi, "Rate-cost tradeoffs in control," *IEEE Trans. Autom. Control*, vol. 64, no. 11, pp. 4525–4540, Nov. 2019.

[33] D. Pollard, "Quantization and the method of k-means," *IEEE Trans. Inf. Theory*, vol. IT-28, pp. 199–205, Mar. 1982.

[34] 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 Trans. Inf. Theory*, vol. 40, no. 6, pp. 1728–1740, Nov. 1994.

[35] T. Linder, "Learning-theoretic methods in vector quantization," in *Principles of Nonparametric Learning*. New York, NY, USA: Springer, 2002, pp. 163–210.

[36] D. Gunduz, P. de Kerret, N. D. Sidiropoulos, D. Gesbert, C. R. Murthy, and M. van der Schaar, "Machine learning in the air," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2184–2199, Oct. 2019.

[37] L. Cregg, F. Alajaji, and S. Yüksel, "Reinforcement learning for zero-delay coding over a noisy channel with feedback," in *Proc. 62nd IEEE Conf. Decis. Control (CDC)*, Dec. 2023, pp. 3939–3944.

[38] L. Cregg, F. Alajaji, and S. Yüksel, "Near-optimality of finite-memory codes and reinforcement learning for zero-delay coding of Markov sources," in *Proc. Amer. Control Conf.*, 2024, pp. 1–11.

[39] H. Permuter, P. Cuff, B. Van Roy, and T. Weissman, "Capacity of the trapdoor channel with feedback," *IEEE Trans. Inf. Theory*, vol. 54, no. 7, pp. 3150–3165, Jul. 2008.

[40] O. Elishco and H. Permuter, "Capacity and coding for the Ising channel with feedback," *IEEE Trans. Inf. Theory*, vol. 60, no. 9, pp. 5138–5149, Sep. 2014.

[41] Z. Aharoni, D. Tsur, Z. Goldfeld, and H. H. Permuter, "Capacity of continuous channels with memory via directed information neural estimator," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 2014–2019, doi: 10.1109/ISIT44484.2020.9174109.

[42] D. Tsur, Z. Aharoni, Z. Goldfeld, and H. Permuter, "Data-driven optimization of directed information over discrete alphabets," *IEEE Trans. Inf. Theory*, vol. 70, no. 3, pp. 1652–1670, Mar. 2024.

[43] A. Gersho and R. Gray, *Vector Quantization and Signal Compression*. New York, NY, USA: Springer, 2012.

[44] Y. Reznik, "An algorithm for quantization of discrete probability distributions," in *Proc. DCC*, 2011, pp. 333–342.

[45] S. Yüksel and T. Linder, "Optimization and convergence of observation channels in stochastic control," *SIAM J. Control Optim.*, vol. 50, no. 2, pp. 864–887, Jan. 2012.

[46] N. Saldi, T. Linder, and S. Yüksel, *Finite Approximations in Discrete-Time Stochastic Control: Quantized Models and Asymptotic Optimality*. Basel, Switzerland: Birkhäuser, 2018.
[47] A. D. Kara and S. Yüksel, "Convergence of finite memory Q learning for POMDPs and near optimality of learned policies under filter stability," *Math. Operations Res.*, vol. 49, pp. 2066–2093, Nov. 2023.
[48] N. Saldi, S. Yüksel, and T. Linder, "On the asymptotic optimality of finite approximations to Markov decision processes with borel spaces," *Math. Oper. Res.*, vol. 42, no. 4, pp. 945–978, Nov. 2017.
[49] C. Szepesvari, "The asymptotic convergence-rate of Q-learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 10, 1998, pp. 1064–1070.
[50] G. Qu and G. Wierman, "Finite-time analysis of asynchronous stochastic approximation and Q-learning," in *Proc. Mach. Learn. Res.*, vol. 125, 2020, pp. 1–21.
[51] P. Chigansky and R. Liptser, "Stability of nonlinear filters in nonmixing case," *Ann. Appl. Probab.*, vol. 14, no. 4, pp. 2038–2056, Nov. 2004.
[52] R. van Handel, "The stability of conditional Markov processes and Markov chains in random environments," *Ann. Probab.*, vol. 37, no. 5, pp. 1876–1925, Sep. 2009.
[53] G. B. Di Masi and L. Stettner, "Ergodicity of hidden Markov models," *Math. Control, Signals, Syst.*, vol. 17, no. 4, pp. 269–296, Oct. 2005.
[54] O. Hernández-Lerma and J. B. Lasserre, *Markov Chains and Invariant Probabilities*. Basel, Switzerland: Birkhäuser-Verlag, 2003.
[55] R. M. Dudley. *Real Analysis and Probability*. 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2002.
[56] S. Yüksel. (2023). *Optimization and Control of Stochastic Systems*. [Online]. Available: https://mast.queensu.ca/~yuksel/LectureNotesOnStochasticOptControl.pdf
[57] R. L. Dobrushin, "Central limit theorem for nonstationary Markov chains. I," *Theory Probab. Appl.*, vol. 1, no. 1, pp. 65–80, Jan. 1956.
[58] M. Hairer. (2010). *Convergence of Markov Processes*. [Online]. Available: https://www.hairer.org/notes/Convergence.pdf
[59] C. Villani, *Optimal Transport: Old and New*. Berlin, Germany: Springer, 2008.
[60] E. Even-Dar and Y. Mansour, "Learning rates for Q-learning," *J. Mach. Learn. Res.*, vol. 5, pp. 1–25, Dec. 2004.
[61] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 129–137, Mar. 1982.
[62] W. Kreitmeier, "Optimal vector quantization in terms of wasserstein distance," *J. Multivariate Anal.*, vol. 102, no. 8, pp. 1225–1239, Sep. 2011.
[63] N. Saldi, S. Yuksel, and T. Linder, "Asymptotic optimality of finite model approximations for partially observed Markov decision processes with discounted cost," *IEEE Trans. Autom. Control*, vol. 65, no. 1, pp. 130–142, Jan. 2020.
[64] J. Kieffer, "Uniqueness of locally optimal quantizer for log-concave density and convex error weighting function," *IEEE Trans. Inf. Theory*, vol. IT-29, no. 1, pp. 42–47, Jan. 1983.
[65] M. Schäl, "Average optimality in dynamic programming with general state space," *Math. Operations Res.*, vol. 18, no. 1, pp. 163–172, Feb. 1993.

Liam Cregg (Graduate Student Member, IEEE) received the B.A.Sc. degree in mathematics and engineering from Queen's University, Canada. He is currently pursuing the M.A.Sc. degree in mathematics and engineering. He will be starting the Ph.D. degree in electrical engineering with ETH Zürich in 2024. His research interests include stochastic control, reinforcement learning, information theory, and probability.

Tamás Linder (Fellow, IEEE) received the M.S. degree in electrical engineering from the Technical University of Budapest, Hungary, in 1988, and the Ph.D. degree in electrical engineering from Hungarian Academy of Sciences in 1992. He was a Post-Doctoral Researcher with the University of Hawaii in 1992 and a Visiting Fulbright Scholar with the Coordinated Science Laboratory, University of Illinois at Urbana–Champaign, from 1993 to 1994. From 1994 to 1998, he was a Faculty Member with the Department of Computer Science and Information Theory, Technical University of Budapest. From 1996 to 1998, he was a Visiting Research Scholar with the Department of Electrical and Computer Engineering, University of California at San Diego. In 1998, he joined Queen's University, where he is currently a Professor of mathematics and engineering with the Department of Mathematics and Statistics. His research interests include communications and information theory, source coding and vector quantization, machine learning, and statistical pattern recognition. He received the Premier's Research Excellence Award of the Province of Ontario in 2002 and the Chancellor's Research Award of Queen's University in 2003. He was an Associate Editor of *Source Coding* and IEEE TRANSACTIONS ON INFORMATION THEORY from 2003 to 2004.

Serdar Yüksel (Senior Member, IEEE) received the B.Sc. degree in electrical and electronics engineering from Bilkent University and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Illinois at Urbana–Champaign in 2003 and 2006, respectively. He was a Post-Doctoral Researcher with Yale University before joining the Department of Mathematics and Statistics, Queen's University, Canada, where is currently a Professor. He is the coauthor of three research books. His research interests include stochastic control, information theory, and probability. He was a recipient of several awards and has been an editor of several journals.