2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Reinforcement Learning for Zero-Delay Coding over a Noisy Channel with Feedback

Liam Cregg, Fady Alajaji, Serdar Yüksel

Abstract— **In Shannon's classical information-theoretic lossy coding problem, one is allowed to encode long sequences of source symbols at once in order to achieve a lower distortion, which is optimal in the limit of unbounded block lengths. Such a block-coding approach is undesirable in many delay-sensitive applications, such as networked control, sensor networks and live-streaming, among others. Accordingly, we are interested in a variant of Shannon's lossy coding problem, where one wishes to send an information source causally at a fixed rate with no delay over a channel with feedback, while minimizing the average distortion at the receiver. Thus, the classical block-coding approach is not viable.**

This problem has previously been studied using stochastic control techniques, leading to existence, structural, and general approximation results. However, these techniques do not provide actual code designs, and they lead to algorithmic implementations that are computationally difficult. To address this problem, we propose a reinforcement learning approach by building on recent results on quantized Q-learning. We consider the case of a finite-alphabet Markov source over a discrete memoryless channel. After developing some supporting technical results on regularity and stability properties of the associated Markov process, we rigorously justify convergence of a quantized Q-learning algorithm to a near-optimal policy for this problem. Finally, we illustrate our theoretical findings via simulations.

I. Introduction

A. Zero-Delay Coding over a Noisy Channel with Feedback

For Shannon's classic lossy coding problem, in which one wishes to compress, transmit, and reconstruct an information source over a noisy channel, one is allowed to encode long sequences of source symbols at once [1]. Such an approach creates a large delay at the encoder, and is thus undesirable for many time-sensitive applications. Accordingly, we consider the zero-delay lossy coding problem, in which one must send an information source causally at a fixed rate over a noisy channel with feedback while minimizing the expected distortion at the receiver.

In the following, we consider a source $\{X_t\}_{t\geq 0}$ that is a time-homogeneous, discrete-time Markov process taking values in a finite set $\mathbb{X}$. We assume that this process is irreducible and aperiodic, and has its transition kernel described by $P(x_{t+1}|x_t)$. We also assume that the distribution of X_0, which we denote by π_0, is available at the encoder and decoder. The channel is a discrete memoryless channel with input and output alphabets $\mathcal{M}$ and $\mathcal{M}'$, respectively, and transition matrix given by $T(q'_t|q_t)$ (where $q_t \in \mathcal{M}$ and $q'_t \in \mathcal{M}'$). Finally, we denote the recontruction sequence as $\{\hat{X}_t\}_{t\geq 0}$, taking values in a finite set $\hat{\mathbb{X}}$. Throughout, we will use the sequence notation $X_{[0,t]} \coloneqq \{X_0, \dots, X_t\}$.

We denote the *encoder policy* by the sequence $\gamma^e = \{\gamma^e_t\}_{t\geq 0}$ and the *decoder policy* by $\gamma^d = \{\gamma^d_t\}_{t\geq 0}$. At time t, we let the encoder have access to all past channel inputs, all past channel outputs (in the form of channel feedback), and all past and present source symbols in order to generate the channel input. That is, $\gamma^e_t : \mathcal{M}^t \times (\mathcal{M}')^t \times \mathbb{X}^{t+1} \to \mathcal{M}$, and $q_t = \gamma^e_t(q_{[0,t-1]}, q'_{[0,t-1]}, X_{[0,t]})$. We call the set of all such encoder policies the set of *admissible encoder policies*, and denote it by Γ^e. Similarly, we allow the decoder to have access to all past and present channel outputs in order to generate the recontruction symbol, so that $\gamma^d_t : (\mathcal{M}')^{t+1} \to \hat{\mathbb{X}}$ and $\hat{X}_t = \gamma^d_t(q'_{[0,t]})$, and we denote these admissible decoder policies by Γ^d. We will denote an admissible joint encoding-decoding policy by $\gamma^{ed} \in \Gamma \coloneqq \Gamma^e \times \Gamma^d$.

For the lossy coding problem, the goal is to minimize the average distortion. In the infinite-horizon case, this is given by

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

where $d : \mathbb{X} \times \hat{\mathbb{X}} \to [0, \infty)$ is a distortion measure and $\mathbf{E}^{\gamma^{ed}}_{\pi_0}$ is the expectation with initial distribution $X_0 \sim \pi_0$ under policy γ^{ed}. We denote the optimal average cost by $J^*(\pi_0) \coloneqq \inf_{\gamma^{ed}\in\Gamma} J(\pi_0, \gamma^{ed})$.

Note that for fixed $q_{[0,t-1]}$, $q'_{[0,t-1]}$ and $X_{[0,t-1]}$, the map $\gamma^e_t(q_{[0,t-1]}, q'_{[0,t-1]}, X_{[0,t-1]}, \cdot)$ is a *quantizer* (i.e., a map from $\mathbb{X}$ to $\mathcal{M}$), which we denote by Q_t. Thus we can view an encoder policy at time t as selecting a quantizer Q_t based on $(q_{[0,t-1]}, q'_{[0,t-1]}, X_{[0,t-1]})$, then quantizing X_t as $q_t = Q_t(X_t)$. Also, since the source alphabet is finite, there clearly exists an optimal decoding policy for every encoding policy. Thus in the following we denote the encoding policy by $\gamma \coloneqq \gamma^e$, and assume a corresponding optimal decoding policy is used. We can then restrict our search to finding optimal encoding policies.

B. Literature Review and Preliminaries

Several important structural results have been established for the above setup. For the finite horizon problem, [2] showed that any encoder policy can be replaced, without performance loss, by one of the form $q_t = \gamma_t(q'_{[0,t-1]}, X_t)$. Furthermore, [3] proved a similar result for an encoder policy

This work was supported by the Natural Sciences and Engineering Research Council of Canada. The first author was also supported by a Queen's University Department of Mathematics and Statistics Summer Research Award.

The authors are with the Department of Mathematics and Statistics, Queen's University, Kingston ON, Canada (liam.cregg@queensu.ca, fa@queensu.ca, yuksel@queensu.ca).

using only the conditional probability $P(X_t \in \cdot | q'_{[0,t-1]})$ and X_t to generate q_t. These results were generalized in [4]–[8].

Stochastic control techniques have been crucial in the study of this problem, sometimes in combination with information-theoretic arguments. For example, [2], [3] use dynamic programming, [8]–[10] use the vanishing discount method, while [7] uses convex analysis. Especially important is [8], which showed the existence of optimal policies for the infinite-horizon problem, which we now review.

Let $\mathcal{P}(\mathbb{X})$ be the space of all probability measures on $\mathbb{X}$, and define $\pi_t \in \mathcal{P}(\mathbb{X})$ as

$$\pi_t(A) \coloneqq P(X_t \in A | q'_{[0,t-1]}).$$

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.

Theorem 1: [8, Theorem 3] There exists a stationary Walrand-Varaiya type policy γ^* that solves the infinite-horizon average cost problem, i.e., one that satisfies

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

For technical reasons, we also introduce the discounted cost problem. In the discounted cost problem, the goal is to minimize, for some $\beta \in (0,1)$,

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

As with the average cost, we denote the optimal discounted cost by $J^*_\beta(\pi_0) \coloneqq \inf_{\gamma \in \Gamma} J_\beta(\pi_0, \gamma)$. Note that the discounted cost problem is not the standard objective from a source-channel coding perspective, but it will be important in our study of the average cost problem. For this reason, we mention the following result on the discounted cost problem.

Proposition 1: [8, Proposition 2] For any $\beta \in (0,1)$, there exists a stationary Walrand-Varaiya type policy γ^* that solves the infinite-horizon discounted cost problem, i.e., one that satisfies

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

If one cannot find an optimal policy for these problems, it may suffice to find a policy that obtains the infimum within some tolerance $\epsilon > 0$, such that $J(\pi_0, \gamma^*) = J^*(\pi_0) + \epsilon$. We will assume that such an ϵ is specified a priori, and so for the rest of the paper we will omit ϵ and simply refer to such a policy as *near-optimal*.

Despite the above existence results, finding an optimal policy (either in closed form or algorithmically) is difficult. Under certain setups for the source and channel, analytical solutions exist. For example, [3] showed memoryless encoding (that is, $q_t = X_t$) is optimal when $\mathbb{X} = \mathcal{M}$ and the channel is symmetric. Furthermore, if the source is independent and identically distributed (i.i.d.), one might attempt a Lloyd-Max style algorithm (in the noisy-channel case, this is often called channel-optimized quantization). See for example [11]–[13] for discussions regarding these types of quantizers. However, even in the i.i.d. case, such algorithms generally rely on necessary (and not sufficient) conditions for optimality and thus may only obtain local rather than global optima. These comparisons will be made explicit using simulations.

Thus, for a general source and channel, finding an optimal encoding policy is an open problem. It is natural then to pursue a reinforcement-learning approach to find a solution. Indeed, [14] used recent results from [15], [16] to rigorously justify convergence of a Q-learning algorithm to a near-optimal solution. However, this was only shown in the case where the channel is noiseless. We generalize this result (and accordingly, the supporting results used in [14]) to the case where the channel is noisy. Additionally, we develop further near-optimality results for the average cost problem and its relation with the discounted cost problem via a coupling argument. Finally, we present numerical studies where we verify the near-optimality of the presented algorithm.

The remainder of the paper is organized as follows. In Section II, we review some stochastic control results relevant to the problem and introduce a proposed Q-learning solution, based on results from [15]. In Section III, we review and generalize some important results from [7], [8], [14] for optimal encoders over a noisy channel with feedback. Section IV contains the final algorithm and a rigorous justification of convergence to a near-optimal policy. Finally, Section V provides some simulation results and a comparison to other encoder policies or algorithms.

II. Q-learning and Quantized Q-learning

Definition 2: We define a *Markov decision process* (MDP) as a 4-tuple $(\mathsf{Z}, \mathsf{U}, P, c)$, where:

1) Z is the *state space*, which we assume is Polish (a Borel subset of a complete, separable metric space).
2) U is the *action space*, also Polish.
3) $P = P(\cdot | z, u)$ is the *transition kernel*, a stochastic kernel on Z given $\mathsf{Z} \times \mathsf{U}$.
4) $c : \mathsf{Z} \times \mathsf{U} \to [0, \infty)$ is the *cost function*.

An admissible MDP policy is a sequence $\widetilde{\gamma} = \{\widetilde{\gamma}_t\}_{t \geq 0}$ such that $\widetilde{\gamma}_t : \mathsf{U}^t \times \mathsf{Z}^{t+1} \to \mathsf{U}$. Such a policy, along with the transition kernel P and an initial distribution $\mu \in \mathcal{P}(\mathsf{Z})$, define a unique distribution on $(\mathsf{Z} \times \mathsf{U})^\infty$ [17, Chapter 2]. We denote the resulting state-action process by $\{Z_t, U_t\}_{t \geq 0}$. The goal for the infinite-horizon, discounted cost case is to find a policy $\widetilde{\gamma}$ minimizing, for some $\beta \in (0,1)$,

$$J_\beta(\mu, \widetilde{\gamma}) \coloneqq \lim_{T \to \infty} E^{\widetilde{\gamma}}_\mu \left[\sum_{t=0}^{T-1} \beta^t c(Z_t, U_t) \right].$$

We define the optimal value function as $J^*_\beta(\mu) \coloneqq \inf_{\widetilde{\gamma}} J_\beta(\mu, \widetilde{\gamma})$. If $\mu = \delta_z$ (that is, a Dirac measure at $z \in \mathsf{Z}$), we denote it by $J^*_\beta(z)$.

A. Q-learning for Finite Models

For finite state and action spaces, a common learning method to find an optimal policy is Q-learning. We define the Q-factor at time $t \geq 0$ as $\mathsf{Q}_t : \mathsf{Z} \times \mathsf{U} \to \mathbb{R}$, and the learning rate as $\alpha_t : \mathsf{Z} \times \mathsf{U} \to \mathbb{R}$. In the Q-learning algorithm, one applies an arbitrary admissible policy $\widetilde{\gamma}$ to select U_t and

collects realizations of the process $\{Z_t, U_t, c(Z_t, U_t)\}_{t\geq 0}$. The Q-factors are then updated according to

$$\mathsf{Q}_{t+1}(z,u) = (1-\alpha_t(z,u))\mathsf{Q}_t(z,u) + \alpha_t(z,u)[c(z,u) + \beta \min_{v\in\mathsf{U}} \mathsf{Q}_t(Z_{t+1},v)].$$

We impose the following assumption on the learning rate α_t.

Assumption 1: For all $(z,u) \in \mathsf{Z}\times\mathsf{U}$ and for all $t \geq 0$,

1) $\alpha_t(z,u) \in [0,1]$.
2) $\alpha_t(z,u) = 0$ unless $(Z_t, U_t) = (z,u)$.
3) $\alpha_t(z,u)$ is a function of $(z_0,u_0),\ldots,(z_t,u_t)$.
4) $\sum_{t\geq 0} \alpha_t(z,u) = \infty$ almost surely.
5) $\sum_{t\geq 0} \alpha_t^2(z,u) < \infty$ almost surely.

Proposition 2: [18] Under Assumption 1, the Q-factors $\{\mathsf{Q}_t\}_{t\geq 0}$ converge almost surely to a limit Q^* such that

$$\tilde{\gamma}^*(z) = \arg\min_{v\in\mathsf{U}} \mathsf{Q}^*(z,v)$$

is an optimal policy (i.e., $J_\beta(\mu,\tilde{\gamma}^*) = J_\beta^*(\mu)$).

Although a powerful algorithm, we cannot apply this to MDPs with continuous state spaces as every state-action pair cannot be visited infinitely often. 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. In [15] and [16], conditions under which the resulting policy is near-optimal for the original MDP are given. We review these next.

B. Q-learning for General Spaces

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

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

We define the maximum radius of the quantizer as

$$d_\infty := \max_{i=1,\ldots,N} \sup_{z\in B_i} \|z - y_i\|.$$

Then the quantized Q-learning algorithm proceeds similarly to the standard Q-learning algorithm: let $\mathsf{Q}_t : \mathsf{Y}\times\mathsf{U} \to \mathbb{R}$ and $\alpha_t : \mathsf{Y}\times\mathsf{U} \to \mathbb{R}$. We apply an arbitrary admissible policy $\widetilde{\gamma}$ to select U_t and collect realizations of the process $\{Y_t, U_t, c(Z_t,U_t)\}_{t\geq 0}$. The Q-factors are then updated according to

$$\mathsf{Q}_{t+1}(y,u) = (1-\alpha_t(y,u))\mathsf{Q}_t(y,u) + \alpha_t(y,u)[c(z,u) + \beta \min_{v\in\mathsf{U}} \mathsf{Q}_t(Y_{t+1},v)]. \quad (1)$$

Assumption 2: Assume that our original MDP has the following properties:

1) The stochastic kernel $P(\cdot|z,u)$ is weakly continuous in (z,u), i.e., $P(\cdot|z_n,u_n) \to P(\cdot|z,u)$ weakly for all $(z_n,u_n) \to (z,u)$.
2) The cost function c is continuous and bounded.
3) The action space U is finite.
4) The state space Z is σ-compact.

We also impose the following properties on α_t and $\widetilde{\gamma}$.

Definition 3: We say a policy $\widetilde{\gamma}$ is a *memoryless exploration policy* if for all $z \in \mathsf{Z}$ and $t \geq 0$,

$$\Pr(\widetilde{\gamma}_t(z) = u_i) = p_i, \quad i = 1,\ldots,|\mathsf{U}|,$$

where $p_i > 0$ for all i and $\sum_i p_i = 1$. That is, the policy chooses actions entirely independently and randomly. For simplicity, we let $p_i = |\mathsf{U}|^{-1}$ and refer to this as a *uniform exploration policy*.

Assumption 3: Assume the following:

1) $\alpha_t(y,u) = \frac{1}{1+\sum_{k=0}^t 1_{(Y_k,U_k)=(y,u)}}$.
2) The policy $\widetilde{\gamma}$ used is a uniform exploration policy.
3) Under $\widetilde{\gamma}$, the state process $\{Z_t\}_{t\geq 0}$ admits a unique invariant measure ψ^*.
4) Under $\widetilde{\gamma}$, every pair $(y,u) \in \mathsf{Y}\times\mathsf{U}$ is visited infinitely often almost surely.

Under the above assumptions, from [15] and [19], we have the following result.

Theorem 2: [15, Theorem 3.2] [19, Theorem 4.27] Under Assumptions 2 and 3, the Q-factors $\{\mathsf{Q}_t\}_{t\geq 0}$ in (1) converge almost surely to a limit Q^*. Furthermore, consider the policy $\tilde{\gamma}^* : \mathsf{Y} \to \mathsf{U}$ given by

$$\tilde{\gamma}^*(y) = \arg\min_{v\in\mathsf{U}} \mathsf{Q}^*(y,v),$$

and extend this policy to Z by making $\tilde{\gamma}^*$ constant over each B_i, $i = 1,\ldots,N$:

$$\tilde{\gamma}^*(z) = \tilde{\gamma}^*(y_i) \quad \text{for all } z \in B_i.$$

For d_∞ close to 0, $\tilde{\gamma}^*$ is near-optimal for the original MDP. To summarize, we can find a near-optimal policy for an MDP with a continuous state space as follows: quantize the state space finely enough, apply the quantized Q-learning algorithm until the Q-factors converge, and extend the resulting policy to the original MDP. In the following sections, we will see how we can use this strategy to solve the zero-delay lossy source-channel coding problem.

III. Zero-Delay Source-Channel Coding as an MDP

Recall the setup from Section I; in particular, the information source X_t, the quantizer Q_t, the channel input q_t, the channel output q'_t, the recontruction symbol $\hat{X}_t$, and the definition of π_t:

$$\pi_t(A) := P(X_t \in A | q'_{[0,t-1]}).$$

Proposition 3: Under a Walrand-Varaiya type policy, the update equation for π_t is given by

$$\pi_{t+1}(x_{t+1}) = \frac{1}{\sum_{q_t} T(q'_t|q_t)\pi_t(Q_t^{-1}(q_t))} \cdot \sum_{q_t} \sum_{x_t\in Q_t^{-1}(q_t)} P(x_{t+1}|x_t)T(q'_t|q_t)\pi_t(x_t), \quad (2)$$

where $Q_t^{-1}(q_t) = \{x \in \mathbb{X} : Q_t(x) = q_t\}$.

Thus, π_{t+1} is conditionally independent of $(\pi_{[0,t-1]}, Q_{[0,t-1]})$ given π_t and Q_t. Then, we have that $\{\pi_t, Q_t\}_{t\geq 0}$ is a controlled Markov chain, and we denote the resulting transition kernel by $P(d\pi_{t+1}|\pi_t, Q_t)$. We wish to define a cost function

for this process in terms of π_t and Q_t that gives the average distortion. The following lemma gives us this cost function.

Lemma 1: For a given Q_t, if an optimal decoder is used, the average distortion is given by

$$c(\pi_t, Q_t) := \sum_{q_t'} \min_{\hat{x}\in\hat{\mathbb{X}}} \Big(\sum_{q_t} \sum_{x_t \in Q_t^{-1}(q_t)} \pi_t(x_t) T(q_t'|q_t) d(x_t, \hat{x}) \Big). \tag{3}$$

By Lemma 1, since we assume we use an optimal decoder for a given Q_t, we have

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

where we recall the notation in Section I for an admissible encoding policy γ, and in Section II for an admissable MDP policy $\widetilde{\gamma}$.

Letting $\widetilde{\mathcal{Q}}$ be the set of all quantizers, we can write the zero-delay lossy source-channel coding problem as an MDP, in the form of Definition 2:

$$(\mathsf{Z}, \mathsf{U}, P, c) = (\mathcal{P}(\mathbb{X}), \widetilde{\mathcal{Q}}, P(\cdot|\pi, Q), c).$$

That is, the state space is the space of probability measures on $\mathbb{X}$, the action space is the set of quantizers, the transition kernel is determined by (2), the cost function is determined by (3), and an admissible MDP policy $\widetilde{\gamma}$ is given by an admissible encoding policy γ.

We use a topology on the set of quantizers $\widetilde{\mathcal{Q}}$ given in [7], [20]. This topology comes from representing a quantizer Q as a stochastic kernel $Q(q|x)$ and considering equivalence classes based on the joint measure $PQ(x,q) = P(x)Q(q|x)$. Under this topology, [7] showed the following:

Lemma 2: [7, Lemma 11]. The transition kernel $P(d\pi_{t+1}|\pi_t, Q_t)$ is weakly continuous in (π_t, Q_t). That is,

$$\int_{\mathcal{P}(\mathbb{X})\times\widetilde{\mathcal{Q}}} f(\pi') P(d\pi'|\pi, Q)$$

is continuous on $\mathcal{P}(\mathbb{X})\times\widetilde{\mathcal{Q}}$ for all continuous bounded f.

At this point, we have Assumptions 2.1-2.4, 3.1, and 3.2. Assumption 3.4 can be shown by taking sufficiently long sequences of (Q_t, q_t', x_t), which are visited infinitely often almost surely due to our assumptions on the source and encoder policy. We next prove that Assumption 3.3 holds for this MDP, i.e., that under a uniform exploration policy on the quantizers, $\{\pi_t\}_{t\geq 0}$ admits a unique invariant measure.

A. Predictor and Filter Stability

To show the desired result, we will need some supporting results from the literature on so-called hidden Markov models. We introduce the *filter*, which is obtained by further conditioning π_t on q_t':

$$\overline{\pi}_t(A) := P(X_t \in A | q'_{[0,t]}).$$

The filter admits a recursion equation similar to (2) (see [21]–[23]). Note that these recursions are dependent on the initialization of π_0, which we call the *prior*. We denote the predictor (respectively, filter) process resulting from the prior $\pi_0 = \nu$ as $\{\pi_t^\nu\}_{t\geq 0}$ (respectively, $\{\overline{\pi}_t^\nu\}_{t\geq 0}$).

Definition 4: Let $A, B \in \mathcal{P}(\mathbb{X})$. We define the total variation norm as

$$\|A - B\|_{TV} := \sup_{\|f\|_\infty \leq 1} \Big| \int f dA - \int f dB \Big|,$$

for f measurable.

Definition 5: We say that the predictor (respectively, filter) process is *stable in total variation in expectation* if for any $\mu, \nu \in \mathcal{P}(\mathbb{X})$ such that μ is absolutely continuous with respect to ν, we have

$$\lim_{n\to\infty} E^\mu[\|\pi_t^\mu - \pi_t^\nu\|_{TV}] = 0.$$

We use the following lemmas to deduce predictor stability.

Lemma 3: [22, Theorem 2.19] The filter is stable in total variation in expectation if and only if the predictor is stable in total variation in expectation.

Lemma 4: [23, Corollary 5.5] Let $\{X_t\}_{t\geq 0}$ be a discrete-time Markov chain and $\{Y_t\}_{t\geq 0}$ be a stochastic process such that the Y_t are conditionally independent given $\{X_t\}_{t\geq 0}$ and $P(Y_t|X_{s\geq 0})$ has the form

$$P(Y_t \in A | X_{s\geq 0}) = \int_A g(X_t, y)\psi(dy),$$

where $g(x,y)$ is a probability density with respect to the σ-finite measure ψ. If g is strictly positive, and $\{X_t\}_{t\geq 0}$ is positive Harris recurrent and aperiodic, then the filter $\overline{\pi}_t(A) := P(X_t \in A | y_{[0,t]})$ is stable in total variation in expectation.

The next lemma follows by applying Lemma 4 and showing that the density $g(x, q')$ is strictly positive.

Lemma 5: Under a uniform exploration policy, the filter process $\{\overline{\pi}_t\}_{t\geq 0}$ is stable in total variation in expectation.

Lemmas 2 and 4 immediately imply the following:

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

The following theorem is inspired by [24, Theorem 3], while using additional Markov properties of $\{\pi_t\}_{t\geq 0}$.

Theorem 3: Under a uniform policy, the process $\{\pi_t\}_{t\geq 0}$ admits a unique invariant measure.

By Theorem 3 we have that Assumption 3.3 is met, and thus the quantized Q-learning results in Section II are applicable.

Remark 1: Controlled Models. Note that many of the above results also hold when the source $\{X\}_{t\geq 0}$ is controlled. That is, a decision maker chooses an action $u_t = f_t(q'_{[0,t]}, u_{[0,t-1]})$ and the source evolves according to $P(x_t|x_{t-1}, u_{t-1})$. See [25] for some generalizations to this setup. However, the optimization is now over both the quantizers and the control actions. Since the quantizer and controller may affect each other's information, the question of dual optimality is more challenging. In the future, we intend to extend our results to this controlled case.

IV. Algorithms

We now present our algorithm to implement quantized Q-learning for the zero-delay coding problem.

A. Quantizing π_t

Since the state space $\mathbb{X}$ is finite, say with $|\mathbb{X}| = m$, then $\mathcal{P}(\mathbb{X})$ is a simplex in $\mathbb{R}^m$. For $n \in \mathbb{N}$ consider the set $\mathcal{P}_n(\mathbb{X}) := \{\hat{\pi} \in \mathcal{P}(\mathbb{X}) : \hat{\pi} = [\frac{k_1}{n}, \ldots, \frac{k_m}{n}], k_i = 0, \ldots, n\}$. For a given π_t, we want to find the nearest $\hat{\pi}_t \in \mathcal{P}_n(\mathbb{X})$. We use an algorithm developed by Reznik [26, Algorithm 1] for this purpose.

Reznik's Algorithm [26, Algorithm 1]

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

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

Recalling the notation of Section II, we have the following.

Lemma 6: [26, Proposition 2] Using Reznik's algorithm, the maximum radius of the quantizer for $\mathcal{P}(\mathbb{X})$ is given by

$$d_\infty = \frac{1}{n}\left(1 - \frac{1}{m}\right).$$

Also note that the number of levels when using Reznik's algorithm is given by $N = \binom{n+m-1}{m-1}$ [26].

B. Quantized Q-learning for Source-Channel Coding

Algorithm 1: Quantized Q-learning for Source-Channel Coding

Require: source alphabet $\mathbb{X}$, transition kernel $P(x_{t+1}|x_t)$, initial distribution π_0, channel kernel $T(q_t'|q_t)$, quantization parameter n, quantizer set $\widetilde{\mathcal{Q}}$, uniform exploration policy γ, discount factor $\beta \in (0, 1)$

Initialize Q-factor $\mathsf{Q}_0 : \mathcal{P}_n(\mathbb{X}) \times \widetilde{\mathcal{Q}} \to \mathbb{R}$
Sample x_0 according to π_0
Quantize π_0 using Reznik's algorithm with parameter n, call this $\hat{\pi}_0$
Select quantizer Q_0 according to γ
$q_0 = Q_0(x_0)$
Generate q_0' according to $T(q_0'|q_0)$
for $t \geq 0$ **do**
  Compute $c(\pi_t, Q_t)$ (see (3))
  Generate x_{t+1} according to $P(x_{t+1}|x_t)$
  Compute π_{t+1} (see (2))
  Quantize π_{t+1} using Reznik's algorithm with parameter n, call this $\hat{\pi}_{t+1}$
  Update Q-factor Q_t (see (1))
  Select quantizer Q_{t+1} according to γ
  $q_{t+1} = Q_{t+1}(x_{t+1})$
  Generate q_{t+1}' according to $T(q_{t+1}'|q_{t+1})$
end for

Theorem 4: In Algorithm 1, the Q-factors $\{\mathsf{Q}_t\}_{t \geq 0}$ converge almost surely to a limit Q^*. Furthermore, if Algorithm 1 is used with large enough n, the encoding policy γ^* given by

$$\gamma^*(\pi) = \arg\min_{Q \in \widetilde{\mathcal{Q}}} \mathsf{Q}^*(\hat{\pi}, Q) \quad \text{for } R(\pi) = \hat{\pi},$$

is near-optimal for the discounted cost problem, where $R(\pi)$ is the output of Reznik's algorithm when the input is π.

Proof: We have shown that Assumption 2 and 3 hold for the zero-delay coding problem. As $n \to \infty$, by Lemma 6, we have $d_\infty \to 0$. Then the result follows by applying Theorem 2. ∎

C. Connection to the Average Cost Problem

Note that we have mostly studied the discounted cost problem, even though the average cost problem is typically the objective in source-channel coding. Here we make explicit the connection to the average cost problem. The following is due to several results in the MDP literature for the average cost problem, see e.g. [27, Theorems 7.3.3-7.3.5].

Lemma 7: Let $\gamma_{\beta,n}$ be the policy obtained by applying Algorithm 1 with parameters β and n. Then there exist $\beta \in (0, 1)$ and $n \in \mathbb{N}$ such that $\gamma_{\beta,n}$ is near-optimal for the average cost problem.

V. Simulations

In our simulations, we use the squared-error distortion measure $d(x, \hat{x}) = (x - \hat{x})^2$. Each data point below is generated by running a finite-horizon ($T = 10^5$) discounted cost problem with a high discount factor ($\beta = 0.9999$).

A. Optimal Quantizers for Symmetric Channels

Here, we consider the case where $\mathbb{X} = \hat{\mathbb{X}} = \mathcal{M} = \{1, \ldots, 4\}$ (that is, we consider 4-level scalar quantizers). Also, $\mathcal{M}' = \{1, \ldots, 5\}$ (the event $q_t' = 5$ can be interpreted as an "erasure" by the channel). The channel is a 4-ary symmetric erasure channel with erasure probability 0.1 and error probability 0.05. The Markov source transition matrix is given by

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

As mentioned in Section I, the case where $\mathbb{X} = \mathcal{M}$ and the channel is symmetric was solved by Walrand and Varaiya [3]. It was shown that encoding is "useless" in this case, i.e., the policy that chooses $Q_t(x_t) = x_t$ is optimal. We see that our algorithm approaches this optimal cost as n increases.

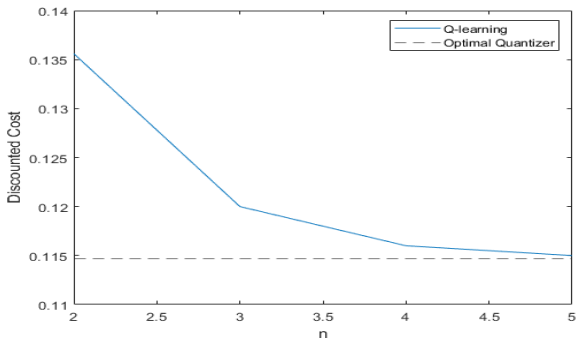

Fig. 1. Comparison with an optimal quantizer.

B. Comparison with a Channel-Optimized Quantizer

A note on optimality: In the following, we consider an i.i.d. source and compare our algorithm to a channel-optimized scalar quantizer (COSQ) (see [11]–[13], [28] for discussions on such quantizers that are robust against channel noise). Note that in the presence of feedback, one may be able to achieve lower distortion with adaptive schemes, e.g., [29], however such a scheme is not exactly zero-delay, since it requires using the channel multiple times to send a single source symbol. Also, such algorithms are only guaranteed to converge to *local* optimality, not necessarily global optimality [13], [28]. On the other hand, running Algorithm 1 for large enough n is guaranteed to give (near) *global* optimality. Of course, for small n in Algorithm 1, the resulting policy may converge to something far from the optimum, and the COSQ may perform better, as we can see below.

In the following, we use the source alphabet $\mathbb{X} = \hat{\mathbb{X}} = \{1, \dots, 4\}$ and the channel input and output are given by $\mathcal{M} = \mathcal{M}' = \{1, 2\}$ (that is, we consider 2-level scalar quantizers). Finally, the channel is given by a binary symmetric channel with error probability 0.1. The source distribution is given by $P = (0.2,\ 0.05,\ 0.4,\ 0.35)$.

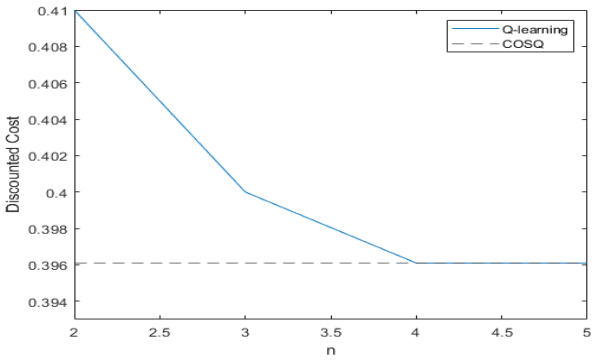

Fig. 2. Comparison with COSQ in i.i.d. case.

References

[1] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*. Urbana, IL: University of Illinois Press, 1949.

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

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

[4] D. Teneketzis, "On the structure of optimal real-time encoders and decoders in noisy communication," *IEEE Trans. Inf. Theory*, vol. 52, pp. 4017–4035, 2006.

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

[6] S. Yüksel, "On optimal causal coding of partially observed Markov sources in single and multi-terminal settings," *IEEE Trans. Inf. Theory*, vol. 59, pp. 424–437, 2013.

[7] T. Linder and S. Yüksel, "On optimal zero-delay coding of vector Markov sources," *IEEE Trans. Inf. Theory*, vol. 60, no. 10, pp. 5975–5991, 2014.

[8] 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, pp. 5968–5980, 2017.

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

[10] 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, pp. 3474–3488, 2022.

[11] N. Farvardin and V. Vaishampayan, "On the performance and complexity of channel-optimized vector quantizers," *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 155–160, 1991.

[12] N. Farvardin, "A study of vector quantization for noisy channels," *IEEE Trans. Inf. Theory*, vol. 36, no. 4, pp. 799–809, 1990.

[13] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. 28, no. 1, pp. 84–95, 1980.

[14] L. Cregg, T. Linder, and S. Yüksel, "Reinforcement learning for the near optimal design of zero-delay codes for markov sources," 2023. [Online]. Available: https://mast.queensu.ca/~yuksel/CLYQ3.pdf

[15] A. Kara, N. Saldi, and S. Yüksel, "Q-learning for MDPs with general spaces: Convergence and near optimality via quantization under weak continuity," *Journal of Machine Learning Research, arXiv:2111.06781*, 2023.

[16] A. Kara and S. Yüksel, "Convergence of finite memory Q-learning for POMDPs and near optimality of learned policies under filter stability," *Mathematics of Operations Research (also arXiv:2103.12158)*, 2023.

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

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

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

[20] S. Yüksel and T. Linder, "Optimization and convergence of observation channels in stochastic control," *SIAM J. on Control and Optimization*, vol. 50, pp. 864–887, 2012.

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

[22] C. McDonald and S. Yüksel, "Stochastic observability and filter stability under several criteria," *IEEE Trans. Autom. Control*, to appear, 2023.

[23] R. van Handel, "The stability of conditional Markov processes and Markov chains in random environments," *Ann. Probab*, vol. 37, pp. 1876–1925, 2009.

[24] L. Stettner, "Ergodic control of partially observed Markov control processes with equivalent transition probabilities," *Applicationes Mathematicae*, vol. 22, pp. 25–38, 1993.

[25] J. C. Walrand and P. Varaiya, "Causal coding and control of Markov chains," *Systems & Control Letters*, vol. 3, pp. 189 – 192, 1983.

[26] Y. Reznik, "An algorithm for quantization of dicrete probability distributions," *DCC 2011*, pp. 333–342, 3 2011.

[27] S. Yüksel, "Optimization and control of stochastic systems," Queen's University, Lecture notes, 2022.

[28] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[29] A. Amanullah and M. Salehi, "Joint source-channel coding in the presence of feedback," in *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*, 1993, pp. 930–934 vol.2.