Reinforcement Learning for Zero-Delay Coding of Markov Sources

Liam Cregg
Mathematics and Statistics
Queen's University
Kingston, Canada
liam.cregg@queensu.ca

Tamás Linder
Mathematics and Statistics
Queen's University
Kingston, Canada
tamas.linder@queensu.ca

Serdar Yüksel
Mathematics and Statistics
Queen's University
Kingston, Canada
yuksel@queensu.ca

Abstract

In the classical lossy coding problem, one is allowed to encode long sequences of source symbols in order to achieve a lower distortion. This is undesirable in many delay-sensitive applications. Accordingly, we consider the zero-delay case where one wishes to encode a source symbol causally. It has been shown that this problem lends itself to stochastic control techniques, leading to existence and structural results. However, attempting to develop analytical solutions or algorithmic implementation using these methods has been computationally difficult. To that end, we propose a reinforcement learning approach. First, we will establish some supporting results on regularity and stability properties. Afterwards, building on recent results on quantized Q-learning, we show that a quantized Q-learning algorithm can be used to obtain a near-optimal policy for this problem. Finally, we provide some relevant simulation results.

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 recontructing the source at the decoder (also without delay). That is, the classical block-coding approach is not viable. We will also allow the encoder to have access to feedback from the channel. Formally, the setup is as follows:

The source $\{X_t\}_{t\geq 0}$ is a time-homogeneous, discrete-time Markov process taking values in a finite set $\mathbb{X}$. Assume that the source is irreducible and aperiodic (and thus admits a unique invariant measure), and has transition kernel $P(x_{t+1}|x_t)$. We also assume that the distribution of X_0, which we denote by π_0, is available at the encoder and decoder. After encoding, the (compressed) information, denoted by q_t, is sent over a discrete noiseless channel with input and output alphabets $\mathcal{M} := \{1, \ldots, M\}$. Thus, the encoder is defined by an encoder policy $\gamma^e = \{\gamma^e_t\}_{t\geq 0}$, where $\gamma^e_t : \mathcal{M}^t \times \mathbb{X}^{t+1} \to \mathcal{M}$, and $q_t = \gamma^e_t(q_{[0,t-1]}, X_{[0,t]})$, where we use the notation $X_{[0,t-1]} = (X_0, \ldots, X_{t-1})$,

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

etc. Then, the decoder generates the reconstruction $\hat{X}_t$ without delay, using decoder policy $\gamma^d = \{\gamma_t^d\}_{t \geq 0}$, where $\gamma_t^d : \mathcal{M}^{t+1} \to \hat{\mathbb{X}}$. Thus we have $\hat{X}_t = \gamma_t^d(q_{[0,t]})$.

Note that for fixed $q_{[0,t-1]}$ and $X_{[0,t-1]}$, the map $\gamma_t^e(q_{[0,t]}, X_{[0,t-1]}, \cdot)$ is a *quantizer* (i.e. a map from $\mathbb{X}$ to $\mathcal{M}$, which we denote by Q_t. So we can view an encoder policy at time t as selecting a quantizer Q_t, then quantizing $q_t = Q_t(X_t)$.

Also, since the source alphabet is finite, there exists an optimal decoding policy for every encoding policy. Thus we will denote the encoding policy by $\gamma := \gamma^e$, and assume it is paired with an optimal decoding policy. We can then restrict our search only to optimal encoding policies.

In general, for the zero-delay coding problem, the goal is to minimize the average-cost/distortion. In the infinite-horizon case, this is given by:

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

where $d : \mathbb{X} \times \hat{\mathbb{X}} \to \mathbb{R}$ is some distortion function and $\mathbf{E}_{\pi_0}^{\gamma}$ is the expectation with $X_0 \sim \pi_0$ and under encoder policy γ. We denote the optimal average cost by $J^*(\pi_0) := \inf_{\gamma \in \Gamma} J(\pi_0, \gamma)$, where Γ is the set of all admissible encoder policies.

B. Literature Review

A number of important structural results have been obtained under this setup. In particular, for the finite horizon problem, [1] 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, [2] 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. These results were generalized in further papers, see e.g. [3]–[7]. Of particular interest for us are the results from [7], which considered existence of optimal policies in the infinite-horizon case. Formally, we introduce the following:

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

$$\pi_t(A) := \Pr(X_t \in A | q_{[0,t-1]})$$

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, i.e. γ_t is a fixed policy.

Theorem 1. *[7, Theorem 3] There exists a stationary Walrand-Varaiya type policy γ^* that solves the infinite-horizon average-cost problem, i.e. one that satisfies:*

$$J(\pi_0, \gamma^*) = J^*(\pi_0)$$

for all π_0.

We note that, despite the key structural results obtained for this problem, finding an optimal policy (either algorithmically or analytically) is difficult. Under certain assumptions on the source and channel, solutions have been found. For example, [2] showed memoryless encoding is optimal when $\mathbb{X} = \mathcal{M}$ and the channel is symmetric, and **Tatikonda2004** studied the Gauss-Markov source when the channel satisfies certain "matching" properties. However, for a general source and channel, finding an optimal encoding policy is an open problem.

We also note the importance of stochastic control approaches in the above structural results. Indeed, under a stochastic control framework, [1], [2] use a dynamic programming recursion in their study of the finite horizon problem, a convex analytic method is employed in [6], and **Tatikonda2009**, [7] study the average-cost optimality equation. In keeping with this strategy, we propose to use a reinforcement learning approach often used to find solutions to stochastic control problems. In particular, we will use results from [8] to rigorously justify convergence of this algorithm to a near-optimal solution.

In our study of the average-cost problem, we will make use of the discounted-cost problem. That is, for some $\beta \in (0,1)$, we wish to minimize:

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

And as with the average cost, we denote the optimal discounted cost by $J_\beta^*(\pi_0) := \inf_{\gamma\in\Gamma} J_\beta(\pi_0,\gamma)$. Note that the discounted-cost problem is in general not very interesting from a source coding perspective. However, we will need it for the rigorous proof of convergence, and we will relate the discounted-cost problem to the average-cost one. For this reason, we also introduce the following result from [7].

Theorem 2. *[7, Proposition 2] There exists a stationary Walrand-Varaiya type policy γ^* that solves the infinite-horizon discounted-cost problem, i.e. one that satisfies:*

$$J_\beta(\pi_0, \gamma^*) = J_\beta^*(\pi_0)$$

for all π_0.

Under certain assumptions, as $\beta \to 1$, the optimal discounted-cost (with an appropriate normalization) approaches the optimal average-cost. We will use Theorem 8 in the Appendix. The assumptions of this theorem have been shown to hold; in particular, [6, Lemma 11] and [7, Lemma 1] give parts (3) and (5), while the rest will follow from our setup of the problem in Section II. By Theorem 8, we have that there exists some sequence $\{\beta_n\}_{n\geq 0}$ with $\beta_n \to 1$ such that:

$$J^*(\pi_0) = \lim_{n\to\infty} (1-\beta_n) J_{\beta_n}^*(\pi_0)$$

Therefore, we may approximate the optimal average cost using the discounted cost. Hence, we will develop an algorithm for the discounted-cost problem which yields a near-optimal policy, and if we do this for large enough β, we obtain a cost close to the optimal average cost.

The remainder of the paper is organized as follows:

- In Section II we formulate the zero delay coding problem as a Markov decision process (MDP) and review key results from [6], [7].
- Section III introduces our proposed Q-learning solution, based on results from [8].

- In Section IV, we prove additional regularity properties of our MDP and relate them to the above Q-learning solution.
- Section V contains the final algorithm and a rigorous justification of convergence to a near-optimal policy for the discounted-cost problem (which is related to the average-cost problem through Theorem 8).
- Finally, Section VI provides some simulation results and a comparison to other encoder policies.

II. Optimal Quantizers

Under Walrand-Varaiya type policies, it was shown in [6] that $\{\pi_t\}$ is a controlled Markov process with control $\{Q_t\}$. More specifically, we have the following result:

Proposition 1. *[6] Under a Walrand-Varaiya type policy, the update equation for π_t is given by*

$$\pi_{t+1}(x_{t+1}) = \frac{1}{\pi_t(Q_t^{-1}(q_t))} \sum_{x_t \in Q_t^{-1}(q_t)} P(x_{t+1}|x_t)\pi_t(x_t) \tag{1}$$

Therefore π_{t+1} is conditionally independent of $(\pi_{[0,t-1]}, Q_{[0,t-1]})$ given π_t and Q_t, and hence $\{\pi_t\}$ is a controlled Markov process with control $\{Q_t\}$.

We will denote the transition kernel induced by the above update equation by $P(d\pi_{t+1}|\pi_t, Q_t)$ (this is a distribution on $\mathcal{P}(\mathbb{X})$). We also define the following cost function for this process in terms of π_t and Q_t (this is the average distortion if the optimal decoder is used for a given Q_t).

$$c(\pi_t, Q_t) := \sum_{i=1}^{M} \min_{\hat{x} \in \hat{\mathbb{X}}} \sum_{x \in Q_t^{-1}(i)} \pi_t(x) d(x, \hat{x}) \tag{2}$$

Note that by this definition of $c(\pi_t, Q_t)$ and our assumption that we are using an optimal decoder for a given encoder of the optimal Walrand-Varaiya type, we have:

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

A. A Topology on Quantizers

As we will be discussing convergence and continuity regarding quantizers, we need to define an appropriate topology. Let $Q : \mathbb{X} \to \mathcal{M}$ be a quantizer, and denote the i^{th} *bin* of Q by $B_i = Q^{-1}(i), i \in \mathcal{M}$. We denote the set of all possible quantizers by $\mathcal{Q}$ (since $\mathbb{X}$ and $\mathcal{M}$ are finite, so is $\mathcal{Q}$). Following [6], [9], we note that Q can alternatively be represented as a stochastic kernel from $\mathbb{X}$ to $\mathcal{M}$ such that $Q(i|x) = 1_{\{x \in B_i\}}, i \in \mathcal{M}$. Then for $P \in \mathcal{P}(\mathbb{X})$, we denote by PQ the joint probability measure $PQ(x,y) = P(x)Q(y|x), (x,y) \in \mathbb{X} \times \mathcal{M}$. If we introduce the equivalence relation $Q \equiv Q'$ iff $PQ = PQ'$, then we can imbue the equivalence classes with the weak convergence topology (that is, we say $Q_n \to Q$ weakly iff $PQ_n \to PQ$ weakly). Under this topology, [6] showed the following property of the controlled Markov chain $\{\pi_t\}$.

Lemma 1. *[6, Lemma 11]. The transition kernel $P(d\pi_{t+1}|\pi_t, Q_t)$ is weakly continuous in (π_t, Q_t). That is,*

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

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

We briefly remark on the setup so far.

Remark 1.

1) We now have a controlled Markov chain $\{\pi_t\}$, which takes values in $\mathcal{P}(\mathbb{X})$, with $\mathcal{Q}$-valued control $\{Q_t\}$. Equipped with the cost function (2), this becomes a Markov decision process (MDP) (see Section III and [10] for more information on MDPs). Theorem 1 tells us that solving the zero delay coding problem is equivalent to finding an optimal policy Walrand-Varaiya type policy for this MDP.

2) We wish to use a reinforcement learning approach to find an optimal or near-optimal policy for this MDP. We note that, although our information source $\mathbb{X}$ is finite, the state space for our new MDP is $\mathcal{P}(\mathbb{X})$ and therefore uncountable. We will see how this impacts reinforcement learning in the following section.

III. Q-LEARNING AND QUANTIZED Q-LEARNING

A. *Q-learning for Finite Models*

As in [10, Chapter 2], we define a *Markov decision process* as a 4-tuple $(\mathsf{X}, \mathsf{U}, P, c)$, where:

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

For now, we will assume X and U are both finite, and deal with the infinite case shortly. An *admissible policy* is a sequence $\gamma = \{\gamma_t\}_{t\geq 0}$ such that $\gamma_t : \mathsf{U}^t \times \mathsf{X}^{t+1} \to \mathsf{U}$. Such a policy, along with the transition kernel P and an initial distribution $X_0 \sim \mu$, define a unique distribution for $(X_t, U_t)_{t\geq 0}$. The goal (for the infinite-horizon, discounted-cost case) is to find a policy γ minimizing:

$$J_\beta(\mu, \gamma) := \lim_{T\to\infty} E_\mu^\gamma \left[\sum_{t=0}^{T-1} \beta^t c(X_t, U_t) \right]$$

for some $\beta \in (0, 1)$.

We define the optimal value function as the above cost when an optimal policy is used:

$$J_\beta^*(\mu) := \inf_\gamma J_\beta(\mu, \gamma)$$

We also note that if $\mu = \delta_x$, we denote the above by $J_\beta^*(x)$. A key result in stochastic control theory is that a function is the optimal value function iff it satisfies the discounted-cost optimality equation (DCOE):

$$J_\beta^*(x) = \min_{u\in\mathsf{U}} \left\{ c(x, u) + \beta \sum_{y\in\mathsf{X}} J_\beta^*(y) P(y|x, u) \right\} \tag{3}$$

A common method to find a policy satisfying the DCOE is Q-learning, which we now define. Here, $\mathsf{Q}_t : \mathsf{X} \times \mathsf{U} \to \mathbb{R}$ is the *Q-factor* at time $t \geq 0$, and $\alpha_t : \mathsf{X} \times \mathsf{U} \to \mathbb{R}$ is the *learning rate*. Suppose that we start at an arbitrary Q_0 and we use some arbitrary admissable policy γ. The Q-factors are updated as follows:

Q-learning

```
Initialize $x_0$ according to $\mu$
for $t \geq 0$ do
    $u_t = \gamma_t(x_t)$
    if $(x, u) = (x_t, u_t)$ then
        $\mathsf{Q}_{t+1}(x,u) = (1-\alpha_t(x,u))\mathsf{Q}_t(x,u) + \alpha_t(x,u)[c(x,u) + \beta \min_{v \in \mathbb{U}} \mathsf{Q}_t(X_{t+1}, v)]$
    else
        $\mathsf{Q}_{t+1}(x,u) = \mathsf{Q}_t(x,u)$
    end if
    Receive $x_{t+1}$ according to $P(\cdot|x_t, u_t)$
end for
```

We impose the following assumption on the learning rate α_t.

Assumption 1. *For all $(x, u) \in \mathsf{X} \times \mathsf{U}$ and for all $t \geq 0$, we have*

1) $\alpha_t(x, u) \in [0, 1]$.
2) $\alpha_t(x, u)$ *is a function of* $(x_0, u_0), \ldots, (x_t, u_t)$.
3) $\sum_{t \geq 0} \alpha_t(x, u) = \infty$.
4) $\sum_{t \geq 0} \alpha_t^2(x, u) < \infty$.

A well-known result in stochastic control theory is the following:

Proposition 2. Watkins *Under Assumption 1, in the above algorithm the Q-factors $\{\mathsf{Q}_t\}_{t \geq 0}$ converge almost surely to:*

$$\mathsf{Q}^*(x,u) := c(x,u) + \beta \sum_{y \in \mathsf{X}} \min_{v \in \mathsf{U}} \{\mathsf{Q}(y, v)\} P(dy|x,u)$$

Note that $\min_{v \in \mathsf{U}} \mathsf{Q}^*(x, v) = J^*_\beta(x)$ and hence the policy made up of these minimizing actions is optimal. Although a powerful algorithm, we cannot apply this to our zero-delay coding problem directly as our state space is infinite, and therefore (3) in Assumption 1 will fail. A potential solution is to use "quantized" Q-learning - that is, we approximate the original MDP using some other MDP with a finite state space, and run Q-learning on this model. Under some additional assumptions, [8] showed that one can indeed achieve near-optimality for the original MDP in this fashion. Note that [8] also considered quantization of the action space, but we will not need this for our application, and so we will assume U is finite.

B. Finite Model Approximations: State Space

Assumption 2. *Our original MDP has the following properties:*

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

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

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

Now let $\psi \in \mathcal{P}(\mathsf{X})$. Then with an abuse of notation we define the resulting conditional distribution

$$\psi(A|y_i) := \frac{\psi(A)}{\psi(B_i)}$$

Let $\hat{\text{MDP}} := (\mathsf{Y}, \mathsf{U}, \hat{c}, \hat{P})$, where $\hat{c}$ and $\hat{P}$ are defined as the mean of the original c and P over the quantization bins. That is:

$$\hat{c}(y_i, u) := \int_{B_i} c(x, u)\psi(dx|y_i)$$
$$\hat{P}(y_j|y_i, u) := \int_{B_i} P(B_j|x, u)\psi(dx|y_i) \tag{4}$$

Then let $\hat{J}_\beta$ be the optimal value function for $\hat{\text{MDP}}$, and note that we can extend this function over X by making it constant over each B_i, i.e.

$$\hat{J}_\beta(x) := \hat{J}_\beta(y_i) \quad \forall x \in B_i$$

We also define:

$$d_\infty := \max_{i=1,\ldots,N-1} \sup_{x,x' \in B_i} ||x - x'||$$

.

And note that under Assumption 2, X is σ-compact, and hence there exists a sequence of partitions $\{B_i\}_{i=1}^N$ such that $d_\infty \to 0$ and $\bigcup_{i=1}^{N-1} B_i \uparrow \mathsf{X}$ as $N \to \infty$. Then we have:

Theorem 3. *[11, Theorem 4.27] Under Assumption 2, we have for all compact $K \subset \mathsf{X}$ and as $d_\infty \to 0$,*

$$\sup_{x_0 \in K} |\hat{J}_\beta(x_0) - J_\beta^*(x_0)| \to 0$$

and

$$\sup_{x_0 \in K} |J_\beta(x_0, \hat{\gamma}) - J_\beta^*(x_0)| \to 0$$

where $\hat{\gamma}$ is the optimal policy of $\hat{\text{MDP}}$ extended to X.

C. Quantized Q-learning

Based on the above theorems, we can find a near-optimal policy for the discounted-cost problem by quantizing our state space finely enough, and finding an optimal policy for $\hat{\text{MDP}}$. However, we do not necessarily know that a Q-learning algorithm for this new MDP will converge, as the quantization process introduces some non-Markovian features into the MDP [8]. A result from [8] does in fact guarantee convergence of a Q-learning algorithm to the optimal Q-factor of $\hat{\text{MDP}}$ under some additional assumptions. The algorithm is exactly as the **Q-learning** algorithm, but with $\mathsf{Q}_t : \mathsf{Y} \times \mathsf{U}$ and $\alpha_t : \mathsf{Y} \times \mathsf{U} \to \mathbb{R}$. That is, line 5 becomes:

$$\mathsf{Q}_{t+1}(f(x), u) = (1 - \alpha_t(f(x), u))\mathsf{Q}_t(f(x), u) + \alpha_t(f(x), u)(c(x, u) + \beta \min_{v \in \mathsf{U}} \mathsf{Q}_t(f(X_{t+1}), v)) \tag{5}$$

Assumption 3. *In the above Q-learning algorithm, we have:*

1)

$$\alpha_t(y,u) = \begin{cases} \frac{1}{1+\sum_{k=0}^{t} 1_{(Y_k,U_k)=(y,u)}} & (Y_t,U_t)=(y,u) \\ 0 & otherwise \end{cases}$$

2) The policy γ chooses control actions independently of everything and randomly, i.e.

$$\Pr(\gamma(\cdot)=u_i)=p_i \quad \forall i=1,\ldots,|\mathsf{U}|$$

where $p_i > 0\,\forall i$ and $\sum_i p_i = 1$.

3) Under the above policy γ, the state process $\{X_t\}_{t\geq 0}$ admits a unique invariant measure ψ^*.

Theorem 4. *[8, Theorem 3.2] Under Assumption 3, for each $(y_i,u)\in\mathsf{Y}\times\mathsf{U}$, $Q_{t+1}(y_i,u)$ in* (5) *converges to:*

$$Q^*(y_i,u)=\hat{c}(y_i,u)+\beta\sum_{y_j\in\mathbb{Y}}\hat{P}(y_j|y_i,u)\min_{v\in\mathbb{U}}Q^*(y_j,v),$$

where $\hat{c}$ and $\hat{P}$ are defined as in (4) *with $\psi=\psi^*$.*

We remark briefly on the connection of this quantized Q-learning to our original zero-delay coding problem:

Remark 2.

1) We were able to reduce the zero-delay coding problem to an MDP in Section II, but the MDP had an infinite state space, and thus it was impossible to use standard Q-learning on it (recall that this MDP had π_t, a probability distribution, as its state).

2) Given the above theorems, we can use a quantized Q-learning algorithm to find a near-optimal policy for this MDP. We just need to confirm that Assumption 2 and Assumption 3 hold. Indeed, from Lemma 1 and the definition of c, we already have that Assumption 2 holds. Parts (1) and (2) of Assumption 3 are determined by algorithm design, and so can be met. We are thus left with proving that part (3) is met, i.e. that the process $\{\pi_t\}_{t\geq 0}$ admits a unique invariant measure under a random exploration policy. The next section is dedicated to showing this fact.

IV. Unique Ergodicity Under a Memoryless Exploration Policy

Recall our setup from Section II. In particular, we have a controlled Markov process $\{\pi_t\}_{t\geq 0}$, with control $\{Q_t\}_{t\geq 0}$, where $Q_t:\mathbb{X}\to\mathcal{M}$. Here $\mathbb{X}$ is our source alphabet, $\mathcal{M}$ is our message set, and we have $Q_t(X_t)=q_t$ as the quantizer output.

We wish to show that if we choose the Q_t randomly, $\{\pi_t\}_{t\geq 0}$ admits a unique invariant measure. Note that under a random exploration policy, we can view a given quantizer output $q_t\in\mathcal{M}$ as an observation of the true state X_t, dependent on an i.i.d. noise variable Z_t. That is, we enumerate the set of quantizers as $\mathcal{Q}=\{Q_1,\ldots,Q_m\}$. Then let $q_t=h(x_t,z_t)=Q_{z_t}(x_t)$, where Z_t is an i.i.d random variable taking values in $\mathcal{Z}:=\{1,\ldots,m\}$, with $Z_0\sim R$ positive everywhere. Then we can use some tools from the literature of Partially Observed Markov Processes (POMPs) to study the ergodicity of $\{\pi_t\}$.

A. Predictor and Filter Merging

Given the above characterization, we can view π_t as what is known as a *predictor* in the POMDP literature. Recall the recursion equation (1) and note that this is dependent on the initialization of π_0, also called the *prior*. We denote the predictor process resulting from the prior $\pi_0 = \nu$ as $\{\pi_t^\nu\}_{t\geq 0}$. A common question in the POMDP literature is when measures such as the predictor are *stable*. More formally, we introduce the following definitions:

Definition 2. *A predictor process $\{\pi_t\}_{t\geq 0}$ is stable in the sense of weak merging in expectation if for any $f \in C_b(\mathbb{X})$ and any prior ν with $\mu \ll \nu$, we have $\lim_{n\to\infty} E^\mu[|\int f d\pi_t^\mu - \int f d\pi_t^\nu|] = 0$.*

Definition 3. *A predictor process $\{\pi_t\}_{t\geq 0}$ is stable in the sense of total variation in expectation if for any prior ν with $\mu \ll \nu$, we have $\lim_{n\to\infty} E^\mu[||\pi_t^\mu - \pi_t^\nu||_{TV}] = 0$. Here $||A-B||_{TV} = \sup_{||f||_\infty \leq 1} |\int f dA - \int f dB|$ for f measurable.*

Note that Definition 3 is stronger than Definition 2. The following result follows from [12, Theorem 2], and making use of the above structure of q_t. The proof can be found in the Appendix.

Theorem 5. *Suppose the following hold:*

1) *There exists a unique invariant measure $\zeta(dx)$ for the Markov process $\{X_t\}_{t\geq 0}$.*
2) *The predictor process $\{\pi_t\}_{t\geq 0}$ is stable in the sense of Definition 2.*
3) *The joint process $\{X_t, \pi_t\}_{t\geq 0}$ is Feller.*

Then the process $\{\pi_t\}_{t\geq 0}$ admits a unique invariant measure.

Note that in our setup in Section I, we assumed $\{X_t\}_{t\geq 0}$ admits a unique invariant measure, so (1) is met. Furthermore, since $\mathbb{X}$ is finite and $\{\pi_t\}_{t\geq 0}$ is Feller by Lemma 1 (see [10, Definition C.3]), we have that $\{X_t, \pi_t\}_{t\geq 0}$ is Feller, so (3) is met. Then we are left with showing (2).

B. Predictor Stability

To show predictor stability, we will utilize results in [13], which relate the stability of the predictor to stability of the *filter*, which we denote by π_t^*. The filter has the same form as the predictor but is further conditioned on q_t (that is, $\pi_t^* = P(x_t | q_{[0,t]})$). We have the following result.

Lemma 2. *[13, Theorem 2.11] The filter merges in total variation in expectation if and only if the predictor merges in total variation in expectation.*

And finally, we show that the filter is stable in total variation in expectation. Recall that we defined $q_t = h(x_t, z_t) = Q_{z_t}(x_t)$, where Q_{z_t} was our random quantizer choice. Then we denote $h_x(\cdot) := h(x, \cdot)$. The following lemma is originally from [14], which was presented in a more general context.

Lemma 3. *[14, Corollary 5.5] Consider the conditional probability measure defined by $R(h_x^{-1}(q))$. If this is strictly positive for all (x, q), then the filter converges in total variation in expectation.*

To show that this density is positive in the quantizer case, recall that under a memoryless exploration policy we have that R is positive everywhere. Then, since $\mathcal{Q}$ contains every posssible quantizer $Q : \mathbb{X} \to \mathcal{M}$, we have that for all (x, q), there exists at least one quantizer such that $Q(x) = q$, and thus $h_x^{-1}(q)$ is nonempty for all (x, q). Then, since R is positive everywhere, we obtain by Lemma 3 that the filter is stable in total variation in expectation.

Note that the above arguments also hold as long as, for all (x, q), our set of quantizers contains *at least one* quantizer Q st $Q(x) = q$. More concretely, we impose the following assumption on our set of quantizers $\mathcal{Q}$.

Assumption 4. *For all (x, q), we have $\{Q \in \mathcal{Q} : Q(x) = q\} \neq \emptyset$.*

To summarize, we have the following theorem.

Theorem 6. *If the set of quantizers $\mathcal{Q}$ satisfies Assumption 4, then under any memoryless exploration policy, the Markov process $\{\pi_t\}_{t\geq 0}$ admits a unique invariant measure.*

Proof. First, under Assumption 4 and by Lemma 3, the filter is stable in total variation in expectation. Then, by Lemma 2, the predictor is stable in total variation in expectation, which implies weak stability in expectation, and so by Theorem 5, we have that $\{\pi_t\}_{t\geq 0}$ admits a unique invariant measure. □

In light of Theorem 6 we have that the quantized Q-learning results in [8] are applicable. We now present some relevant algorithms.

V. Algorithms

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 a given π_t and n, we wish to find the nearest (in terms of Eucilidean distance) $\hat{\pi}_t = [\frac{k_1}{n}, \ldots, \frac{k_m}{n}]$, where $k_i \in \mathbb{Z}_{ge0}$. Then we can use the algorithm in e.g. [15], [16] to quantize π_t as follows.

Algorithm 1: Predictor Quantization

```
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 $\delta_i$ 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 $(\frac{k'_1}{n}, \ldots, \frac{k'_m}{n})$
```

We have the following lemma regarding the radius of these quantization bins under the above algorithm.

Lemma 4. *[15, Proposition 2] The maximum radius of the quantization regions for π_t under the L_∞ norm is given by*

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

Also note that the number of bins for π_t when using **Algorithm 1** is related to n by the following relation: # bins $= \binom{n+m-1}{m-1}$ [15].

B. Quantized Q-learning

Using the above algorithm to quantize π_t, we have the following algorithm for quantized Q-learning.

Algorithm 2: Quantized Q-learning

```
Require: source alphabet $\mathbb{X}$, transition kernel $P(x_{t+1}|x_t)$, initial distribution $\pi_0$, quantization parameter $n$, quantizer set $\mathcal{Q}$, exploration policy $\gamma$, time horizon $T$
Initialize Q-table of size $\binom{n+m-1}{m-1} \times |\mathcal{Q}|$
Initialize $x_0$ according to $\pi_0$
Quantize $\pi_0$ using Algorithm 1, call this $\hat{\pi}_0$
Select quantizer $Q_0$ according to $\gamma$
$q_0 = Q_0(x_0)$
for $t = 0$ to $T - 1$ do
    Compute $c(\pi_t, Q_t)$ (see (2))
    Receive $x_{t+1}$ according to $P(x_{t+1}|x_t)$
    Receive $\pi_{t+1}$ according to update equation (see (1))
    Quantize $\pi_{t+1}$ using Algorithm 1, call this $\hat{\pi}_{t+1}$
    Update Q-table (see (5))
    Select quantizer $Q_{t+1}$ according to $\gamma$
    $q_{t+1} = Q_{t+1}(x_{t+1})$
end for
return $\gamma^*(\pi) = \text{argmin}_{Q \in \mathcal{Q}}(\text{Q-table}(\pi, Q))$
```

Theorem 7. *Under Assumption 4 and as $n \to \infty$, the above algorithm gives a near-optimal policy for the discounted-cost, infinite-horizon zero-delay coding problem.*

Proof. First, Assumption 4 allows us to use Theorem 6 to guarantee a unique invariant measure for $\{\pi_t\}$ under a random exploration policy. Therefore Assumption 3 is met. Then from Theorem 4, this algorithm converges. Furthermore, as $n \to \infty$, by Lemma 4 we get $d_\infty \to 0$. Then by Theorem 3, the policy we get from this algorithm performs near-optimally when applied to the to the original MDP for the discounted-cost, infinite-horizon problem. □

VI. Examples

For all the following, we will use MSE as our distortion measure, i.e. $d(x, \hat{x}) = (x - \hat{x})^2$, and a high discount factor (0.9999). We approximate the long-term discounted-cost by calculating a finite-horizon discounted-cost with $T = 10^6$.

A. *Effects of Increasing n for a Fixed Rate*

Say we consider all the two-cell quantizers (i.e. $\mathcal{M} := \{1, 2\}$) on $\mathbb{X} = \{1, \ldots, 5\}$ (i.e. $m = 5$) with a randomly generated transition matrix. We let n in the quantization of π_t vary, and run **Algorithm 2** for these different values of n. Recall that the number of bins in this quantization is given by $\binom{n+m-1}{m-1}$. The first graph below shows the performance gain as we increase the number of bins, while the second shows the convergence of the algorithm for different numbers of bins. Each line in the second graph corresponds to increasing the value of n by 1.

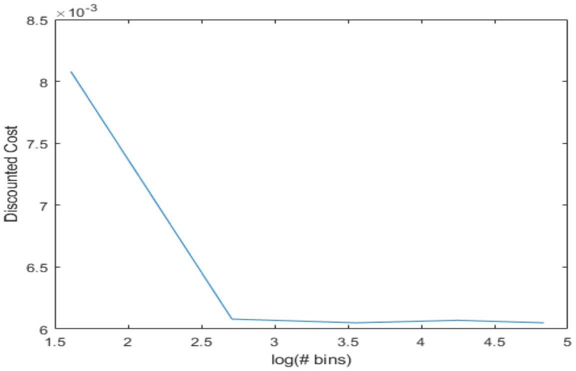

Fig. 1. Long-term discounted-cost of learned policies

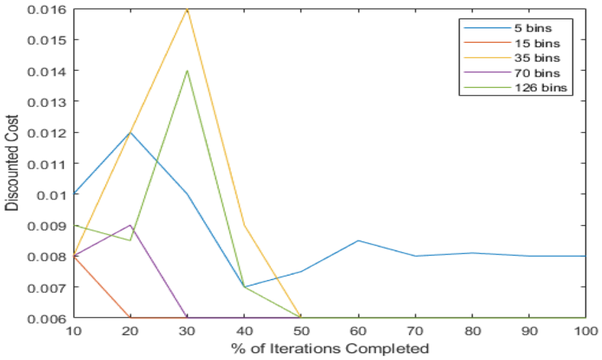

Fig. 2. Convergece of learned policies

Note that the quantization gains are not significant after $n = 2$ (which corresponds to 15 bins), which indicates that this is a sufficient quantization level for a near-optimal policy.

Similarly, for 2-cell quantizers on $\mathbb{X} = \{1, \ldots, 16\}$,

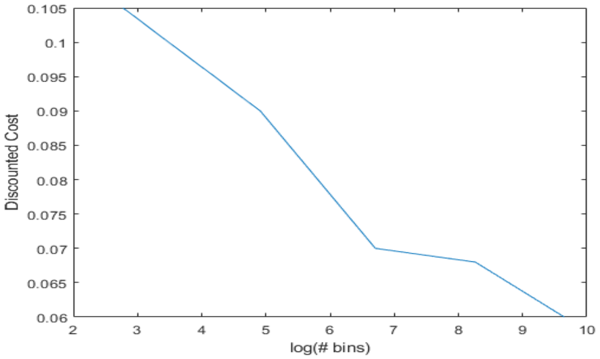

Fig. 3. Long-term discounted-cost of learned policies

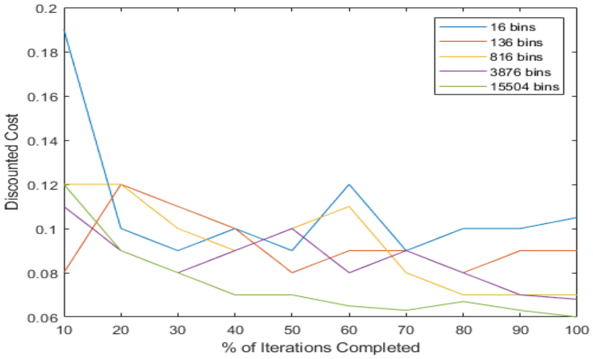

Fig. 4. Convergece of learned policies

In this example, additional quantization may be required to get closer to the optimal policy.

B. Comparison to Lloyd-Max Quantizer

In these simulations, we plot the distortion for different values of n and for different quantizer rates (i.e. different sizes of $\mathcal{M}$). We also plot the comparison with a Lloyd-Max quantizer (another common algorithm to find an optimal quantizer), and note that our algorithm results in a lower discounted-cost. We use the source alphabet $\mathbb{X} = \{1, \ldots, 13\}$, and the rate is calculated by $\log_2(|\mathcal{M}|)$.

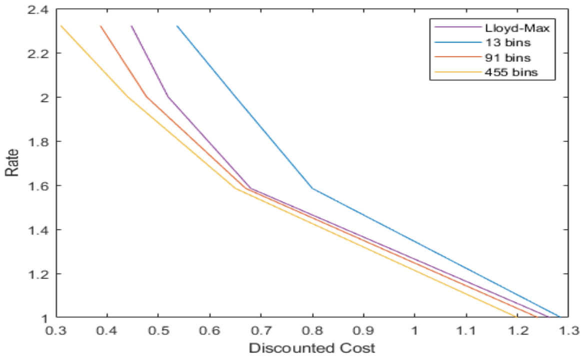

Fig. 5. Comparison with Lloyd-Max in Markov case

Note that for coarse quantizations, a Lloyd-Max quantizer may perform better, but as we increase the quantization, the Q-learning approach gives a lower distortion for a given rate.

Finally, we also note that our algorithm matches very closely with a Lloyd-Max quantizer in the case where the source is i.i.d., as can be seen below. Note that in this case, the Q-learning algorithm only visits one state (the one given by quantizing the distribution of X_0), and hence raising n provides no performance gain. Here we use $\mathbb{X} = \{1, \ldots, 8\}$.

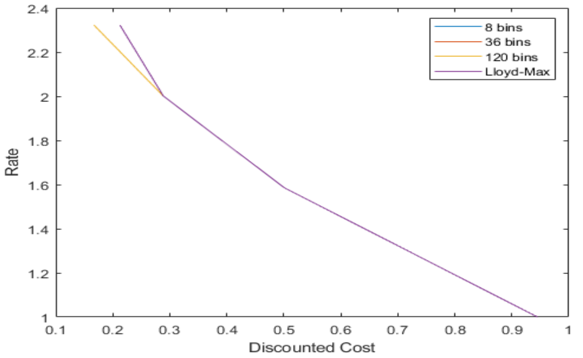

Fig. 6. Comparison with Lloyd-Max in i.i.d. case

We now remark on the algorithm performance and feasibility:

Remark 3.

1) As previously mentioned, the number of bins grows quickly with n, and so running the algorithm with high n requires high amounts of memory in order to store a large Q-table. However, the number of actual visited $\hat{\pi}_t$ tends to be much lower than the total number of bins, and so it may be possible to increase the efficiency of this algorithm by quantizing π_t in a non-uniform fashion (the results from [8] used for convergence of the algorithm allow a non-uniform quantization).
2) As the number of bins grows, the time for the algorithm to converge becomes much higher, since it must visit more states. One must trade much longer algorithm runtime for potential performance increases. Also, since

we only have weak continuity of the transition kernel for $\{\pi_t\}$, we do not have a bound on what quantization level we need to obtain a given performance (results from [8] provide these bounds under stronger notions of continuity). Therefore it is somewhat trial-and-error to find the appropriate quantization level for a given application.

3) As mentioned in Section II, $\{\pi_t\}$ updates in a Markovian fashion, and so we do not need to store the previous values of π_t. Futhermore, from Section IV we have that the process $\{\pi_t\}$ is stable (i.e. it "forgets" an incorrect prior) under the random exploration policy used in learning. Therefore we do not need to worry about errors accumulating over time, and in fact we can start the algorithm from any distribution π_0 (the choice to initialize to the stationary distribution in **Algorithm 2** is just for convenience).

VII. Future Work

We would like to extend the algorithm to the case when $\mathbb{X}$ is continuous. The primary obstable here would be that the number of possible quantizers is infinite, and so a method of quantizing the space of quantizers would need to be developed. We note however that quantizing the action space (in this context, the set of quantizers) is considered in the results from [8], and also all of the predictor stability/unique invariance arguments in Section IV hold when $\mathbb{X}$ is continuous. So with some minor alterations, the algorithm will still converge in the continuous case.

Finally, we wish to consider the case when the channel over which q_t is sent is noisy. While most of the results from Section II carry over in this case, additional considerations will have to be taken in other sections (e.g. in proving uniqueness of the invariant measure).

Appendix

Theorem 8. NearOptimalityQuantized *Consider an MDP* $(\mathsf{X}, \mathsf{U}, P, c)$*, and suppose the following hold:*

1) c is continuous, nonnegative, and bounded.

2) U is bounded.

3) $P(\cdot|x,u)$ is weakly continuous in (x,u).

4) X is compact.

5) The family of functions $\{h_\beta : \beta \in (0,1)\}$, where

$$h_\beta(x) := J^*_\beta(x) - J^*_\beta(x_0)$$

for some fixed $x_0 \in \mathsf{X}$, is uniformly bounded and equicontinuous.

Then there exist a constant $g^* \geq 0$ and a measurable function $f^* : \mathsf{X} \to \mathsf{U}$ such that the policy $\gamma^* = \{f^*\}_{t\geq 0}$ is optimal for the average-cost problem and g^* is the optimal value function, i.e.

$$g^* = J^*(x) = J(x, \gamma^*)$$

Furthermore, $g^* = \lim_{n\to\infty}(1-\beta_n)J^*_{\beta_n}(x)$ for some sequence $\{\beta_n\}_{n\geq 0}$ such that $\beta_n \to 1$.

Proof of Theorem 5

Proof. Throughout, we use the notation $\nu(f) := \int f d\nu$. Assume that $m_1, m_2 \in \mathcal{P}(\mathbb{X} \times \mathcal{P}(\mathbb{X}))$ are two invariant measures for the joint process $\{X_t, \pi_t^\nu\}$. Then their projections on $\mathbb{X}$ are invariant for $\{X_t\}_{t\geq 0}$. Then, by unique invariance of $\zeta(dx)$ we have

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

Then we show that $m_1(F) = m_2(F)$ for each F on a set of measure-determining functions [12], namely those s.t. $F(x,\nu) = \phi(x)H(\nu(\phi_1), \ldots, \nu(\phi_l))$, 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 $l \in \mathbb{N}$.

Let S be the transition operator associated with the process $\{X_t, \pi_t\}$. Then by invariance we have for $i = 1, 2$:

$$m_i(F) = \int\limits_{\mathbb{X}\times\mathcal{P}(\mathbb{X})} \frac{1}{n}\sum_{j=0}^{n-1} S^j F(x,\nu) P_{m_i}(d\nu|x)\zeta(dx)$$

And thus,

$$\begin{aligned}
&|m_1(F) - m_2(F)| \\
&\leq \int\limits_{\mathbb{X}\times\mathcal{P}(\mathbb{X})\times\mathcal{P}(\mathbb{X})} \frac{1}{n}\sum_{j=0}^{n-1} |S^j F(x,\nu_1) - S^j F(x,\nu_2)| \\
&\qquad\qquad\qquad\qquad \cdot P_{m_1}(x,\nu_1)P_{m_2}(x,v_2)\zeta(dx) \\
&\leq L_H||\phi|| \int\limits_{\mathbb{X}\times\mathcal{P}(\mathbb{X})\times\mathcal{P}(\mathbb{X})} \frac{1}{n}\sum_{j=0}^{n-1} E^\mu[\sum_{i=1}^{l} |\pi_j^{\nu_1}(\phi_i) - \pi_j^{\nu_2}(\phi_i)|] \\
&\qquad\qquad\qquad\qquad \cdot P_{m_1}(x,\nu_1)P_{m_2}(x,v_2)\zeta(dx)
\end{aligned}$$

Since the predictors are stable in the sense of Definition 2, and by the dominated convergence theorem, the last line converges to zero as $n \to \infty$. □

Note that the above result concerns the joint process $\{X_t, \pi_t\}_{t\geq 0}$. The following theorem from [12] extends this to $\{\pi_t\}_{t\geq 0}$.

Theorem 9. *[12, Theorem 3] If the joint process $\{X_t, \pi_t\}_{t\geq 0}$ is Feller and admits at most one invariant measure, then $\{\pi_t\}_{t\geq 0}$ admits at most one invariant measure.*

Note that the above Feller assumption is trivially satisfied in the case where $\mathbb{X}$ is finite, since $\{\pi_t\}_{t\geq 0}$ is Feller by Lemma 1.

References

[1] H. Witsenhausen, "On the structure of real-time source coders," *Bell System Technical Journal*, vol. 58, no. 6, pp. 1437–1451, 1979.

[2] J. Walrand and P. Varaiya, "Optimal causal coding-decoding problems," *IEEE Trans. Inf. Theory*, vol. IT-29, no. 6, pp. 814–820, Nov. 1983.

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

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

[5] S. Yüksel, “On optimal causal coding of partially observed Markov sources in single and multi-terminal settings,” *IEEE Transactions on Information Theory*, vol. 59, pp. 424–437, Jan. 2013.

[6] 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, Oct. 2014.

[7] R. 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.

[8] A. Kara, N. Saldi, and S. Yüksel, “Q-learning for MDPs with general spaces: Convergence and near optimality via quantization under weak continuity,” 2021. DOI: 10.48550/ARXIV.2111.06781.

[9] 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.

[10] O. Hernandez-Lerma and J. Lasserre, *Discrete-time Markov Control Processes*. Springer, 1996.

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

[12] G. D. Masi and Ł. Stettner, “Ergodicity of hidden markov models,” *Math. Control Signals Syst.*, vol. 17, no. 4, pp. 269–296, Oct. 2005.

[13] C. McDonald and S. Yuksel, “Converse results on filter stability criteria and stochastic non-linear observability,” 2018. DOI: 10.48550/ARXIV.1812.01772.

[14] 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.

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

[16] N. Saldi, S. Yüksel, and T. Linder, “Asymptotic optimality of finite model approximations for partially observed markov decision processes with discounted cost,” *IEEE Transactions on Automatic Control*, vol. 65, no. 1, pp. 130–142, 2020.