Quantized Q-Learning for Near-Optimal Quantizers

Liam Cregg

July 19, 2022

1 Zero-Delay Lossy Coding and Optimal Quantizers

We are interested in a variant of Shannon's lossy source coding problem: Given an information source $\{X_t\}_{t\geq 0}$ from a finite alphabet $\mathbb{X}$, we wish to compress this source at rate R bits per source symbol, and then reproduce the source as $\{\hat{X}_t\}_{t\geq 0}$, where $\hat{\mathbb{X}}$ is the (also finite) reproduction alphabet. In particular, a $(2^{RT}, T)$ block code encodes T source symbols $X_{[0,T-1]}$ as $\eta^T : \mathbb{X}^T \to \{1, \ldots, 2^{RT}\}$, and decodes them as $\gamma^T : \{1, \ldots, 2^{RT}\} \to \hat{\mathbb{X}}^T$. The goal is often to minimize the distortion, given by

$$D_T(R) := \frac{1}{T} E[\sum_{t=0}^{T-1} d(X_t, \hat{X}_t)]$$

We call a rate-distortion pair (R, D) achievable if there exists a sequence of $(2^{RT}, T)$ codes (η^T, γ^T) such that

$$\limsup_{T\to\infty} D_T(R) \leq D$$

The minimum achievable distortion for a given rate R is denoted by $D(R)$, and classically we can obtain this distortion by taking the block length T to infinity (provided that the source is stationary and ergodic), i.e.

$$D(R) = \lim_{T\to\infty} D_T(R)$$

However, this approach requires encoding very large blocks of data at a time, and hence incurs large delay. This delay may not be allowable in many real-time applications, and so we tackle a variant of this lossy coding problem, where we enforce zero delay (hence a block coding approach is not viable). We will assume throughout that the source $\{X_t\}_{t\geq 0}$ is a discrete-time Markov process with probability matrix P, which is irreducible and aperiodic (and thus admits a unique invariant measure, denoted by π_0). After encoding, the (compressed) information is sent over a discrete noiseless channel with input and output alphabets $\mathcal{M} := \{1, \ldots, M\}$.

Thus, the encoder is defined by a quantization policy $\Pi = \{\eta_t\}_{t\geq 0}$, where $\eta_t : \mathcal{M}^t \times \mathbb{X}^{t+1} \to \mathcal{M}$. That is, the encoder can use all past quantization outputs and all past and current source inputs to generate the current quantization output. This can be viewed as the quantization policy selecting a quantizer $Q_t : \mathbb{X} \to \mathcal{M}$ using past information, then quantizing X_t as $q_t = Q_t(X_t)$ [5]. Then, the decoder generates the reconstruction $\hat{X}_t$ without delay, using decoder policy $\gamma = \{\gamma_t\}_{t\geq 0}$, where $\gamma_t : \mathcal{M}^{t+1} \to \hat{\mathbb{X}}$. Thus we have $\hat{X}_t = \gamma_t(q_{[0,t]})$. Note that, since the source alphabet is finite, there exists an optimal decoding policy for every encoding policy. Thus, we restrict our search to finding an optimal encoding policy and assume the corresponding optimal decoding policy is used.

In general, for the zero-delay coding problem, the goal is to minimize the average cost/distortion. In the infinite horizon case with cost function/distortion measure $c : \mathbb{X} \times \hat{\mathbb{X}} \to [0, \infty)$, this is given by:

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

However, for the time being we will consider the discounted cost problem, as this problem is easier to tackle using Q-learning methods (to be discussed later). Thus, for some $\beta \in (0, 1)$, we wish to minimize:

$$J_\beta(\pi_0, \Pi, \gamma) := \lim_{T\to\infty} \mathbf{E}_{\pi_0}^{\Pi,\gamma} \left[\frac{1}{T} \sum_{t=0}^{T-1} \beta^t c(X_t, U_t) \right]$$

For the finite horizon problem, policies using only the conditional probability measure $\pi_t = P(x_{t+1}|q_{[0,t-1]})$ and x_t to generate q_t have been shown to be optimal by Walrand and Varaiya [11]. That is, for every admissable policy, there exists a policy of the form $Q_t = \eta_t(\pi_t)$ and $q_t = Q_t(X_t)$ that performs at least as well. Such policies are called Walrand-Varaiya type or Markov policies, and denoted by Π_W. If η_t does not depend on t, we call such policies stationary and denote the set of these policies by Π_{WS}. In [12], Walrand and Varaiya's result was also shown to apply to the infinite horizon discounted cost problem, and in fact that the optimal policy is stationary (that is, there exists an optimal (deterministic) quantization policy in Π_{WS} for the infinite horizon discounted cost problem).

Importantly, it was shown in [12] that 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 $\{Q_t\}$. Using conditional probability properties, we obtain the update equation for π_t:

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

Therefore, in theory one could use dynamic programming principles to run an iteration algorithm on this controlled Markov process in order to obtain the optimal policy $\Pi \in \Pi_{WS}$. However, in practice this proves to be difficult (see e.g. [12]), and hence we propose to use Q-learning in order to find the optimal quantization policy. To this end, we will utilize the recent work of [1] in the near-optimality of policies obtained through Q-learning under quantization. First, we will need some results on the properties of the controlled Markov process $\{\pi_t\}$.

1.1 A Topology on Quantizers

As we will be discussing convergence and continuity regarding quantizers, we need to define an appropriate topology. Viewing a quantizer Q as a map from $\mathbb{X}$ to $\mathcal{M}$, we denote the i^{th} *bin* of Q as $B_i = Q^{-1}(i), i = 1, \ldots, M$, and we denote the set of all possible quantizers by $\mathcal{Q}$. Following [5], we note that a quantizer with bins $\{B_1, \ldots, B_M\}$ can alternatively be represented as a stochastic kernel from $\mathbb{X}$ to M such that $Q(i|x) = 1_{x \in B_i}, i = 1, \ldots, M$. Then if P is a probability measure on $\mathbb{X}$, we denote by PQ the joint probability measure $PQ(x, y) = P(x)Q(y|x)$. If we introduce the equivalence relation $Q \equiv Q'$ iff $PQ = PQ'$, then we can imbue these equivalence classes with the weak convergence topology (that is, we say $Q_n \to Q$ iff $PQ_n \to PQ$). Under this topology, [5] showed the following lemma for the controlled Markov chain $\{\pi_t\}$.

Lemma 1.1. *[5, 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 and bounded f.

Proof. The proof can be found in its entirety in [5], but essentially follows by writing the integral as a sum over $\mathcal{M}$, and using the additional fact that $\pi_n Q_n \to \pi Q$ in total variation. □

2 Unique Ergodicity of $\{\pi_t\}$

A desirable feature of $\{\pi_t\}$ (and one that we will need later for Q-learning) is that under a random exploration policy Π^*, the process admits a unique invariant measure. To prove this, note that under

a randomized policy, we can view a given quantizer output q_t as an observation Y_t (living in $\mathbb{Y} := \mathcal{M}$) of the true state X_t, dependent on an i.i.d noise variable Z_t. More concretely, let $m := |\mathcal{Q}|$ (m is finite since $\mathbb{X}$ is finite). Then enumerate the quantizers in $\mathcal{Q}$ from 1 to m, i.e. $\mathcal{Q} = \{Q_1, \dots, Q_m\}$. Finally, let $y_t := q_t = h(x_t, z_t) = Q_{z_t}(x_t)$, where Z_t is an i.i.d random variable taking values in $\mathcal{Z} := \{1, \dots, m\}$. Then we can use some tools from the literature of Partially Observed Markov Processes (POMPs) to determing the ergodicity of $\{\pi_t\}$.

2.1 Predictor and Filter Merging

For a fixed $x \in \mathbb{X}$, we denote $h(x, \cdot) := h_x(\cdot) : \mathcal{Z} \to \mathbb{Y}$, and let the noise process $\{Z_t\}_{t\geq 0}$ have probability measure $Z_0 \sim R$. Since the noise process is assumed independent of the state process $\{X_t\}_{t\geq 0}$, we can write the following update equation for the joint process $\{X_t, Y_t\}_{t\geq 0}$

$$P((X_{t+1}, Y_{t+1})|(X, Y)_{[0,t]} = (x, y)_{[0,t]}) = R(h_{x+1}^{-1}(y_{t+1}))P(x_{t+1}|x_t)$$

It follows that $\{X_t, Y_t\}_{t\geq 0}$ is Markov, with a probability measure on $\mathbb{X}^{\mathbb{Z}_{\geq 0}} \times \mathbb{Y}^{\mathbb{Z}_{\geq 0}}$, endowed with the product topology. Since this measure depends on the distribution of X_0, we denote this measure by P^μ where $X_0 \sim \mu$.

Given the above characterization, we can view π_t as what is known as a *predictor* in the POMDP literature (that is, $\pi_t(x_t) = P(x_t|y_{[0,t-1]})$). 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 ν as $\{\pi_t^\nu\}_{t\geq 0}$. A common question in the POMDP literature is when measures such as the predictor are *stable*, which essentially means that the process eventually forgets an incorrect prior, i.e. $\nu \neq \mu$. More formally, we introduce the following definitions

Definition 2.1. *Two sequences of probability measures $\{P_t\}_{t\geq 0}$ and $\{Q_t\}_{t\geq 0}$ merge weakly if $\forall f \in C_b(\mathbb{X})$, we have $\lim_{t\to\infty} |\int f dP_t - \int f dQ_t| = 0$.*

Definition 2.2. *For two probability measures P and Q, the total variation norm is given by $||P - Q||_{TV} = \sup_{||f||_\infty \leq 1} |\int f dP - \int f dQ|$ for f measurable.*

Definition 2.3. *A predictor process 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 2.4. *A predictor process is stable in the sense of total variation in expectation if for any prior ν with $\mu \ll \nu$, we have $\lim_{n\to\infty} E^\mu[||\pi_t^\mu - \pi_t^\nu||_{TV}] = 0$.*

Note that merging (respectively, stability) in total variation implies merging (respectively, stability) weakly since $C_b(\mathbb{X})$ is a subset of measurable and bounded functions.

The above definitions are of interest due to a result from [6, Theorem 2] relating stability to ergodicity. We state and prove a slightly modified version of this result here, adapted for our setup.

Theorem 2.1. *Assume that there exists a unique invariant measure $\zeta(dx, dy)$ for the Markov process $\{X_t, Y_t\}_{t\geq 0}$ and that the predictor process $\{\pi_t^\mu\}_{t\geq 0}$ is stable in the sense of 2.3. Then there is at most one invariant measure for the the joint Markov process $\{X_t, Y_t, \pi_t^\nu\}$ for any prior ν.*

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

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

Then we show that $m_1(F) = m_2(F)$ for each F on a set of measure-determining functions [6], namely those s.t. $F(x, y, \nu) = \phi(x, y)H(\nu(\phi_1), \ldots, \nu(\phi_l))$, where $\phi \in C(\mathbb{X} \times \mathbb{Y}), \phi_1, \ldots, \phi_l \in C(\mathbb{X}), H$ is bounded and Lipschitz continuous with constant L_H, and $l \in \mathbb{N}$.

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

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

And thus,

$$\begin{aligned}
&|m_1(F) - m_2(F)| \\
\leq &\int_{\mathbb{X}\times\mathbb{Y}\times\mathcal{P}(\mathbb{X})\times\mathcal{P}(\mathbb{X})} \frac{1}{n} \sum_{j=0}^{n-1} |S^j F(x, y, \nu_1) - S^j F(x, y, \nu_2)| P_{m_1}(x, y, \nu_1) P_{m_2}(x, y, v_2)\zeta(dx, dy) \\
\leq &L_H ||\phi|| \int_{\mathbb{X}\times\mathbb{Y}\times\mathcal{P}(\mathbb{X})\times\mathcal{P}(\mathbb{X})} \frac{1}{n} \sum_{j=0}^{n-1} E_{xy}[\sum_{i=1}^{l} |\pi_j^{\nu_1}(\phi_i) - \pi_j^{\nu_2}(\phi_i)|] P_{m_1}(x, y, \nu_1) P_{m_2}(x, y, v_2)\zeta(dx, dy)
\end{aligned}$$

Since the predictors are stable in the sense of 2.3, and by the dominated convergence theorem, the last line converges to zero as $n \to \infty$. □

It can be shown through a tightness argument (see [6, Theorem 3]) that the above also implies that the Markov process $\{Y_t, \pi_t^\nu\}$ has at most one invariant measure.

2.2 Predictor Stability

In light of 2.1, we want to show that our predictor process $\{\pi_t\}_{t\geq 0}$ is stable in the sense of 2.3. While stability (and even unique ergodicity) of the predictor have been studied extensively (see e.g. [9], [2] for discussions), the assumptions of these theorems are in general quite strong, and usually not applicable for large classes of sources. For example, they may require that the state has a uniformly positive transition density, etc. [9].

In order to generalize our sources and quantizers as much as possible, we will utilize results in [7], which relate the stability of the predictor to stability of the *filter*. The filter has the same form as the predictor but is further conditioned on y_t (that is, $\pi_t^* = P(x_t|y_{[0,t]})$). We have the following result.

Lemma 2.2. *[7, Theorem 2.11] The filter merges in total variation in expectation if and only if the predictor merges in total variation in expectation.*

Proof. As shown in [7], one can write equivalent conditions for stability (in TV in expectation) in terms of an expectation conditioned on the intersection of sigma fields. These intersections are the same for the predictor and the filter, so the result follows. □

And finally, a result to prove that the filter is stable in total variation in expectation,

Lemma 2.3. *[3, Corollary 5.5] Recall the conditional probability measure defined by $R(h_x^{-1}(y))$. If this is strictly positive for all x,y, then the filter converges in total variation in expectation.*

To show that this density is positive in the quantizer case, consider the case when R is positive everywhere. Then, since $\mathcal{Q}$ contains every posssible quantizer $Q : \mathbb{X} \to \mathcal{M}$, we have that $\forall (x, y)$, there exists at least one quantizer s.t. $Q(x) = y$, and thus $h_x^{-1}(y)$ has positive measure $\forall (x, y)$. Then, since R is positive everywhere, we obtain by 2.3 that the filter is stable in total variation in expectation.

Therefore, by 2.2, the predictor is stable in total variation in expectation, which implies weak stability in expectation, and so by reftheorem:unique we have that $\{\pi_t, Y_t\}_{t\geq 0}$ admits at most one invariant measure.

Note that the above arguments hold as long as, $\forall (x, y)$, our set of quantizers contains *at least one* quantizer Q st $Q(x) = y$, and therefore we could operate with a reduced set of quantizers. If we further impose that, there are an *equal* number of quantizers mapping x to each $y \in \mathcal{M}$, then we have that the process $\{Y_t\}_{t\geq 0}$ is described entirely by R and is independent of $\{\pi_t\}_{t\geq 0}$. Then, the fact that $\{\pi_t, Y_t\}_{t\geq 0}$ admits at most one invariant measure implies that $\{\pi_t\}$ admits at most one invariant measure. More concretely, we impose the following assumption on our set of quantizers $\mathcal{Q}$.

Assumption 2.1. *Let $\bar{Q}_{xy} = \{Q \in \mathcal{Q} : Q(x) = y\}$. Then $\mathcal{Q}$ satisfies the following properties:*

(i) $\forall x \in \mathbb{X}$ and $\forall y \in \mathcal{M}$ we have $|\bar{Q}_{xy}| \geq 1$.

(ii) $\forall x \in \mathbb{X}$ and $\forall y_1, y_2 \in \mathcal{M}$, we have $|\bar{Q}_{xy_1}| = |\bar{Q}_{xy_2}|$.

Finally, the existence of an invariant measure for $\{\pi_t\}$ is guaranteed since $\{\pi_t\}$ is weak Feller (see 1.1) on a compact space $\mathcal{P}(\mathbb{X})$. Therefore we have the following lemma.

Lemma 2.4. *If the set of quantizers $\mathcal{Q}$ satisfies 2.1, the Markov process $\{\pi_t\}_{t\geq 0}$ admits a unique invariant probability measure.*

In light of all these properties of the process $\{\pi_t\}$, we now consider the following recent work on quantized Q-learning.

3 Near-Optimality of Quantized Policies

By viewing quantization as a measurement kernel and using recent results on convergence of Q-learning algorithms for POMDPs, [1] showed that under mild conditions (namely, weak continuity of the transition kernel and unique ergodicity of the state process), quantized Q-learning (that is, Q-learning where the action and state spaces are quantized versions of the original MDP) leads to asymptotically optimal policies as the number of quantization bins increases. In particular, for any compact $K \subset \mathbb{X}$,

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

where $\hat{J}_\beta$ is the optimal value function for the finite model obtained by quantizing the state and action spaces. This value function (and the policy yielding this value function) can then be extended to the original MDP by making it constant over the quantization bins.

We can obtain such a near-optimal policy by running the "standard" Q-learning algorithm but viewing our quantized state as the true state, i.e.

$$\begin{aligned} Q_{t+1}(q(x), u) = (1 - \alpha_t(q(x), u)) Q_t(q(x), u) + \\ \alpha_t(q(x), u)(c(x, u) + \beta \min_{v \in \mathbb{U}} Q_t(q(X_{t+1}), v)) \end{aligned} \tag{2}$$

Note that the above Q and q are different from those used in Section 1, but are used here for consistency with [1].

We note that, given Lemmas 1.1 and 2.4, the desired properties of $\{\pi_t\}$ hold. Thus, quantized Q-learning is applicable to the zero-delay lossy coding problem. In summary, we will quantize the state and actions of this controlled Markov chain (that is, we will quantize π_t and Q_t) and run Q-learning on this finite model approximation to obtain a near-optimal quantization policy $\hat{\Pi}$ for the original zero-delay coding problem.

4 Algorithms

4.1 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 belief π_t and "granularity" 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 $k_i \in \mathbb{Z}$. We will denote the set of these "types" (as they are referred to in information theory contexts) by $\mathbb{Z}_n$. Then we can use the algorithm in e.g. [10], [8] to quantize the predictor, as follows.

Algorithm 1: Predictor Quantization

Set Parameters
  Predictor ($\pi_t = [p_1, \ldots, p_m]$)
  Granularity (n)
for $i = 1, \ldots, m$
  $k'_i = \lfloor np_i + \frac{1}{2} \rfloor$
end
Set $n' = \sum_i k'_i$
if $n = n'$
  Return $[\frac{k'_1}{n}, \ldots, \frac{k'_m}{n}]$
else
  for $i = 1, \ldots, m$
    $\delta_i = k'_i - np_i$
  end
  Sort δ_i s.t. $\delta_{i_1} \leq \ldots \leq \delta_{i_m}$
  Set $\Delta = n' - n$
  if $\Delta > 0$
    $$k_{i_j} = \begin{cases} k'_{i_j} & j = 1, \ldots, m - \Delta \\ k'_{i_j} - 1 & j = m - \Delta + 1, \ldots, m \end{cases}$$
  else
    $$k_{i_j} = \begin{cases} k'_{i_j} & j = 1, \ldots, |\Delta| \\ k'_{i_j} - 1 & j = |\Delta| + 1, \ldots, m \end{cases}$$
  Return $[\frac{k'_1}{n}, \ldots, \frac{k'_m}{n}]$

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

Lemma 4.1. *[10, Proposition 2] The maximum radius of the quantization regions for π_t under the L_∞ norm is given by*

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

4.2 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

Set Parameters
  State space ($\mathbb{X}$) and state transition kernel ($P(x_{t+1}|x_t)$)
  Distribution of X_0 (π_0)
  Granularity of types for quantization of π_t (n)
  Set of quantizers ($\mathcal{Q}$)
Initialize arbitrary Q-table of size $|Q_n| \times |\mathcal{Q}|$ (see [10] for definition of Q_n)
Initialize state x_0 according to π_0
Quantize π_0 according to 4.1, call this $\hat{\pi}_0$
Select quantizer Q_0 according to randomized exploration policy Π^*
Quantize x_0 according to $q_0 = Q_0(x_0)$
Decode q_0 optimally, $\hat{x}_0 = \gamma_0(q_0)$
for $t = 0 \dots T-1$
  Receive cost of quantization according to $(x_t - \hat{x}_t)^2$
  Receive x_{t+1} according to $P(x_{t+1}|x_t)$
  Receive π_{t+1} according to (1)
  Quantize π_{t+1} according to 4.1, call this $\hat{\pi}_{t+1}$
  Update Q-table using (2)
  Select quantizer Q_{t+1} according to randomized exploration policy Π^*
  Quantize x_{t+1} according to $q_{t+1} = Q_{t+1}(x_{t+1})$
  Decode q_t optimally, $\hat{x}_t = \gamma_t(q_t)$
end

4.2.1 Proof of Convergence

The proof that this algorithm converges to the optimal policy as n grows follows directly from [1, Theorem 3.2]. Under a randomized (and independent) exploration policy Π^* (e.g. as described in 2), we have the following theorem for convergence.

Theorem 4.2. *Assume that the set of quantizers $\mathcal{Q}$ satisfies 2.1. Then the algorithm in 4.2 converges to*

$$Q^*(\hat{\pi}_i, Q) = C^*(\hat{\pi}_i, Q) + \beta \sum_{\hat{\pi}_j \in Q_n} P^*(\hat{\pi}_j|\hat{\pi}_i, Q) \min_{v \in Q} Q^*(\hat{\pi}_j, v)$$

Here, P^* and C^* are defined by

$$C^*(\hat{\pi}_i, Q) = \int_{B_i} c(\pi, Q)\phi_i(d\pi)$$

and

$$P^*(\hat{\pi}_j|\hat{\pi}_i, Q) = \int_{B_i} P(B_j|\pi, Q)\phi_i(dx)$$

where

$$\phi_i(A) := \frac{\phi_{\Pi^*}(A)}{\phi_{\Pi^*}(B_i)}, \quad \forall A \subset B_i, \quad \forall i \in \{1, \dots, |\mathbb{Z}_n|\}$$

and ϕ_{Π^*} is the (unique) invariant measure of $\{\pi_t\}_{t\geq 0}$ under the exploration policy Π^*. Also, $B_i \subset \mathcal{P}(\mathbb{X})$ is the bin under the quantization algorithm in 4.1 (i.e. $\{B_i\}_{i=1}^{|\mathbb{Z}_n|}$ is a partition of $\mathcal{P}(\mathbb{X})$).

Furthermore, the policy $\bar{\Pi}(\hat{\pi}) = \operatorname{argmin}_{Q \in \mathcal{Q}} Q^*(\hat{\pi}, Q)$ *satisfies*

$$\sup_{\pi \in \mathcal{P}(\mathbb{X})} |\hat{J}_\beta(\pi, \bar{\Pi}) - J^*_\beta(x_0)| \to 0$$

as $n \to \infty$.

Proof. This is essentially due to [1, Theorem 3.2 and Corollary 3.3]. From Lemmas 1.1 and 2.4, we have that all of the necessary assumptions are met, and due to Lemma 4.1, we have that the maximum radius of the quantization bins $b_\infty \to 0$ as $n \to \infty$. The result follows. □

5 Examples

5.1 Trivial Example

We start with the simplest case, where $\mathbb{X} = \{1, 2\}$. Let the transition matrix of $\{X_t\}_{t \geq 0}$ be $T = \begin{pmatrix} 0.75 & 0.25 \\ 0.75 & 0.25 \end{pmatrix}$. The only useful quantizers to pick in this situation are constant ones, that is $Q_i(x) = i \quad \forall x \in \mathbb{X}, i = 1, 2$.

Now suppose we wish to quantize π_t very coarsely, say $n = 2$. Then we have $Z_n = \{(1, 0), (\frac{1}{2}, \frac{1}{2}), (0, 1)\}$, so our Q-table is of size 3×2.

Running the above algorithm for $T = 10^5$ iterations gives the following Q-table (the Q-table was initialized to 10).

	Quantizer	
State	4.77	5.27
	10	10
	10	10

Note that in this case, it doesn't matter what n is, since π_t is constant. Nevertheless, the algorithm converges to the correct optimal quantizer (it should always quantize to 1 since the value 1 appears more frequently).

5.2 More Interesting Examples

Say we consider all the two-cell quantizers on $\mathbb{X} = \{1, \ldots, 5\}$, with a randomly generated transition matrix. We run the above algorithm with varying quantization levels. Note that the number of bins is related to n by the following relation: # bins $= \binom{n+m-1}{m-1}$ (recall that $m = |\mathbb{X}|$) [10]. We get the following graphs relating the long-term discounted cost to the quantization level.

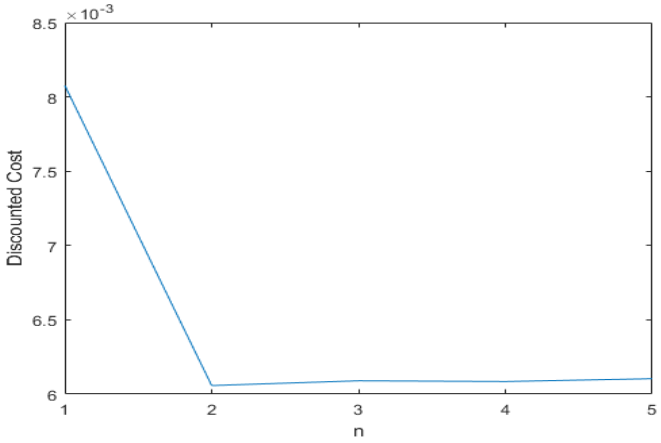

Figure 1: Long-term discounted cost of learned policies

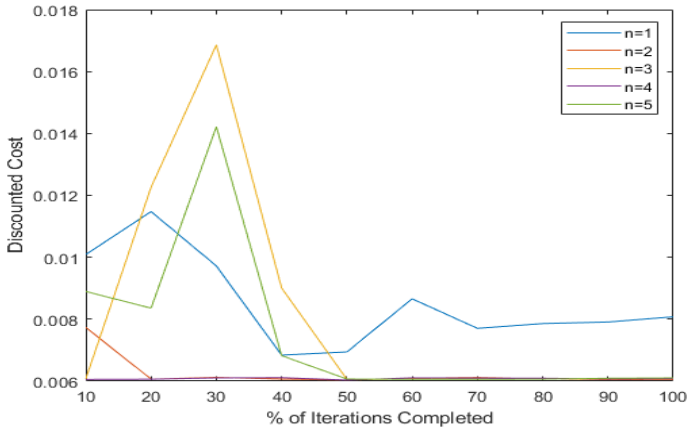

Figure 2: Convergece of learned policies

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

Similarly, we have the following graphs for $\mathbb{X} = \{1, \ldots, 16\}$.

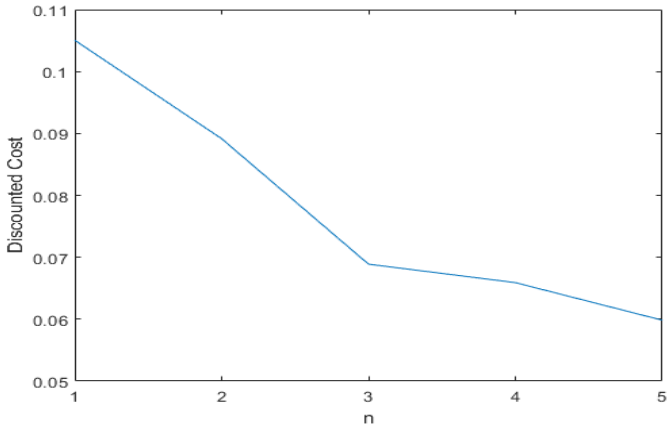

Figure 3: Long-term discounted cost of learned policies

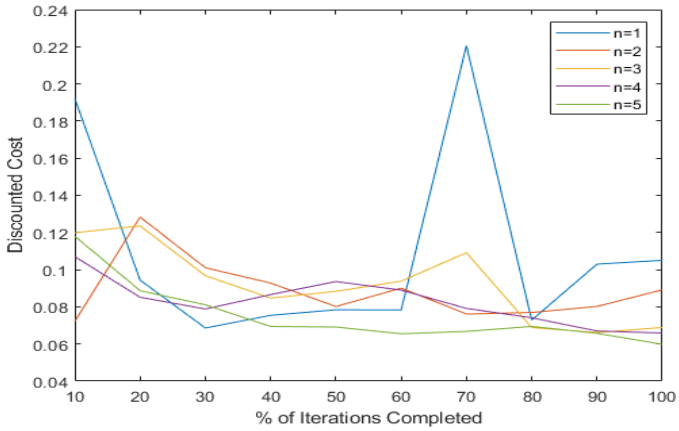

Figure 4: Convergece of learned policies

Here, additional quantization may be required to get closer to the optimal policy. Unfortunately, the number of bins grows quickly with n, and so further increasing the granularity becomes very computationally demanding. However, the number of actual visited states 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 restricting the state space of π_t.

6 Future Work

We note that the above Q-learning algorithm only converges when $\beta \in (0, 1)$ (i.e. for the discounted cost problem) and therefore is not applicable for the average cost problem, which is generally what we are interested in for source coding. It may be possible to adapt the Q-learning algorithm to solve

the average cost problem (see e.g. [4]), but this would require an extension of the results in [1] to this algorithm, which may require additional constraints on the source $\mathbb{X}$.

Furthermore, 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 [1], so convergence should follow with some minor alterations.

References

[1] N. Saldi A.D. Kara 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. URL: https://arxiv.org/abs/2111.06781.

[2] R. Douc and C. Matias. "Asymptotics of the Maximum Likelihood Estimator for General Hidden Markov Models". In: *Bernoulli* 7.3 (2001), pp. 381–420. ISSN: 13507265. URL: http://www.jstor.org/stable/3318493 (visited on 07/13/2022).

[3] R. van Handel. "The stability of conditional Markov processes and Markov chains in random environments". In: *The Annals of Probability* 37.5 (Sept. 2009). DOI: 10.1214/08-aop448. URL: https://doi.org/10.1214%2F08-aop448.

[4] D. Bertsekas J. Abounadi and B. V. S. "Learning algorithms for Markov decision processes with average cost". English (US). In: *SIAM Journal on Control and Optimization* 40.3 (2002), pp. 681–698. ISSN: 0363-0129. DOI: 10.1137/S0363012999361974.

[5] T. Linder and S. Yüksel. "On optimal zero-delay coding of vector Markov sources". In: *IEEE Trans. Inf. Theory* 60.10 (Oct. 2014), pp. 5975–5991.

[6] G.B. Di Masi and Ł. Stettner. "Ergodicity of Hidden Markov Models". In: *Math. Control Signals Syst.* 17.4 (Oct. 2005), pp. 269–296. DOI: 10.1007/s00498-005-0153-8.

[7] C. McDonald and S. Yuksel. *Converse Results on Filter Stability Criteria and Stochastic Non-Linear Observability*. 2018. DOI: 10.48550/ARXIV.1812.01772. URL: https://arxiv.org/abs/1812.01772.

[8] S. Yüksel N. Saldi and T. Linder. "Asymptotic Optimality of Finite Model Approximations for Partially Observed Markov Decision Processes With Discounted Cost". In: *IEEE Transactions on Automatic Control* 65.1 (2020), pp. 130–142. DOI: 10.1109/TAC.2019.2907172.

[9] R. Liptser P. Chigansky and R. van Handel. "Intrinsic methods in filter stability". In: *Handbook of Nonlinear Filtering* (Aug. 2009).

[10] Y.A. Reznik. "An algorithm for quantization of dicrete probability distributions". In: *DCC 2011* (Mar. 2011), pp. 333–342.

[11] J. Walrand and P. Varaiya. "Optimal causal coding-decoding problems". In: *IEEE Trans. Inf. Theory* IT-29.6 (Nov. 1983), pp. 814–820.

[12] R. Wood, T. Linder, and S. Yüksel. *Optimal Zero Delay Coding of Markov Sources: Stationary and Finite Memory Codes*. 2016. DOI: 10.48550/ARXIV.1606.09135. URL: https://arxiv.org/abs/1606.09135.