Control Theory Seminar

Liam Cregg

January 11, 2023

1 Zero-Delay Lossy Coding and Optimal Quantizers

1.1 Setup

- Information source $\{X_t\}_{t\geq 0}$, taking values in $\mathbb{X}$. Time-homogeneous, and in general can be $\mathbb{R}^d$-valued (but for some proofs we'll only prove the finite-alphabet case). Markov with initial distribution π_0 and transition kernel $T(dx_{t+1}|x_t)$. Has a unique invariant distribution.
- Reproduction alphabet, $\hat{\mathbb{X}}$, finite.
- In general, encode T source symbols as $\eta^T : \mathbb{X}^T \to \{1, \ldots, 2^{RT}\}$, send through noiseless channel, and decode them as $\gamma^T : \{1, \ldots, 2^{RT}\} \to \hat{\mathbb{X}}^T$ (these are measurable). We will call the message set $\mathcal{M} := \{1, \ldots, M\}$.
- Expected distortion
$$D_T(R) := \frac{1}{T} E[\sum_{t=0}^{T-1} d(X_t, \hat{X}_t)]$$
where d is the (additive) distortion measure.
- Classically, we can achive min distortion by letting $T \to \infty$, but interested in case where $T = 1$.

So how do we set this up as a MDP?

- Quantization policy $\Pi = \{\eta_t\}_{t\geq 0}$ where $\eta_t : \mathcal{M}^t \times \mathbb{X}^{t+1} \to \mathcal{M}$, decoding policy $\gamma = \{\gamma_t\}$ where $\gamma_t : \mathcal{M}^{t+1} \to \hat{\mathbb{X}}$. I.e. $q_t = \eta_t(q_{[0,t-1]}, x_{[0,t]}), \hat{(x)}_t = \gamma_t(q_{[0,t]})$.
- We wish to minimize (for the finite horizon problem)
$$\mathbf{E}^{\Pi}_{\pi_0}\left[\frac{1}{T}\sum_{t=0}^{T-1} d(X_t, \hat{X}_t)\right]$$
or for infinite horizon
$$J(\pi_0, \Pi) := \limsup_{T\to\infty} \mathbf{E}^{\Pi}_{\pi_0}\left[\frac{1}{T}\sum_{t=0}^{T-1} d(X_t, \hat{X}_t)\right]$$
- But we will also consider the discounted cost problems
$$\mathbf{E}^{\Pi}_{\pi_0}\left[\frac{1}{T}\sum_{t=0}^{T-1} \beta^t d(X_t, \hat{X}_t)\right]$$
and
$$J_\beta(\pi_0, \Pi) := \lim_{T\to\infty} \mathbf{E}^{\Pi}_{\pi_0}\left[\frac{1}{T}\sum_{t=0}^{T-1} \beta^t d(X_t, \hat{X}_t)\right]$$

Theorem 1.1. *For the finite horizon problem, there exists an optimal quantization policy of the form $\{\eta_t\}$ where η_t uses only $q_{[0,t-1]}, x_t$ to generate q_t. i.e. $q_t = \eta_t(q_{[0,t-1]}, x_t)$.*

Proof. Fix a decoding policy $\{\gamma_t\}$. Define the process $v_t = (q_{[0,t-1]}, x_t)$. Then $\{v_t\}$ is Markov given q_t, i.e. $P(v_t|v_{[0,t-1]}, q_{[0,t-1]}) = P(v_t|v_{t-1}, q_{t-1})$. Then by total expectation, the finite horizon cost becomes

$$\begin{aligned}
&\mathbf{E}^{\Pi}_{\pi_0}\left[\frac{1}{T}\sum_{t=0}^{T-1}\mathbf{E}[d(x_t,\gamma_t(q_{[0,t]}))|q_{[0,t-1]},x_{[0,t]},q_t]\right]\\
=&\mathbf{E}^{\Pi}_{\pi_0}\left[\frac{1}{T}\sum_{t=0}^{T-1}F_t(x_t,q_{[0,t-1]},q_t)\right]\\
=&\mathbf{E}^{\Pi}_{\pi_0}\left[\frac{1}{T}\sum_{t=0}^{T-1}F_t(v_t,q_t)\right]
\end{aligned}$$

By MDP results (viewing v_t as state, q_t as action), there exists an optimal control law of the form $\phi_t(v_t)$. □

Define $\pi_t(A) = P(x_t \in A|q_{[0,t-1]}) \in \mathcal{P}(\mathbb{X})$.

Theorem 1.2. *For the finite horizon problem, there exists an optimal quantization policy of the form $\{\eta_t\}$ where η_t uses only π_t, x_t to generate q_t. i.e. $q_t = \eta_t(\pi_t, x_t)$.*

Mention memory spaces of quantizers.
These results were extended to the infinite horizon case and the average cost case in future work.

1.2 (π_t, Q_t) as a controlled Markov chain

- In light of Thm 1.2, we can think of this optimal quantizer policy as one that uses π_t to select a quantizer $Q_t : \mathbb{X} \to \mathcal{M}$, then encodes x_t as $q_t = Q_t(x_t)$.

Theorem 1.3. *The process (π_t, Q_t) is a controlled Markov chain (with state π_t and action Q_t). That is, π_{t+1} is conditionally independent of $(\pi_{[0,t-1]}, Q_{[0,t-1]})$ given (π_t, Q_t).*

Proof. We have

$$\begin{aligned}
\pi_{t+1}(dx_{t+1}) &= \frac{P(dx_{t+1}, q_t|q_{[0,t-1]})}{P(q_t|q_{[0,t-1]})}\\
&= \frac{\int_{x_t}\pi_t(dx_t)P(q_t|\pi_t,x_t)T(dx_{t+1}|x_t)}{\int_{x_{t+1}}\int_{x_t}\pi_t(dx_t)P(q_t|\pi_t,x_t)T(dx_{t+1}|x_t)}\\
&= \frac{1}{\pi_t(Q_t^{-1}(q_t))}\int_{Q_t^{-1}(q_t)}T(dx_{t+1}|x_t)\pi_t(dx_t)
\end{aligned}$$

□

For this controlled Markov chain, our cost can be written as

$$c(\pi_t, Q_t) := \sum_{i=0}^{M}\inf_{\hat{x}\in\hat{\mathbb{X}}}\int_{Q_t^{-1}}\pi_t(x)d(x,\hat{x}) \tag{1}$$

1.3 Topology on Quantizers

- We denote the i^{th} *bin* of Q as $B_i = Q^{-1}(i), i = 1, \ldots, M$
- Can alternatively be represented as a stochastic kernel from $\mathbb{X}$ to M such that $Q(i|x) = 1_{x \in B_i}, i = 1, \ldots, M$
- Denote by PQ the joint probability measure $PQ(x, y) = P(x)Q(y|x)$. Use 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, we have the following

Theorem 1.4. *The transition kernel $T(d\pi_{t+1}|\pi_t, Q_t)$ is weakly continuous in (π_t, Q_t). That is,*

$$\int_{\mathcal{P}(\mathbb{X}) \times \mathcal{Q}} f(\pi^{'}) T(d\pi^{'}|\pi, Q)$$

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

1.4 Motivation for Q-learning

Given the above setup, we can obtain strong structural results. For example, the existence of Walrand-Varaiya type stationary policies, for both finite and infinte horizon and in both the discounted and average cost cases. Also have existence of finite memory codes that are near-optimal.

Also, one might be tempted to run a dynamic programming recursion to try to solve this problem, but in practice this gets computationally intractable after a few iterations. So rather we will attempt to use Q-learning, and in particular, quantized Q-learning.

1.5 Q-learning

We will temporarily use different notation to align with standard Q-learning literature. We will consider the infinite horizon, discounted cost problem.

- Assume finite state and action spaces $\mathbb{X}$ and $\mathbb{U}$, cost function $c : \mathbb{X} \times \mathbb{U} \to \mathbb{R}$, transition kernel $T(x_{t+1}|x_t, u_t)$.
- We want a policy $\gamma = \{\gamma_t\}$ where $\gamma_t : \mathbb{X}^t + 1 \times \mathbb{U}^t \to \mathbb{U}$
- Define the optimal value function

$$J^*_\beta(x_0) = \inf_\gamma J_\beta(x_0, \gamma)$$

- The optimal value function satisfies the DCOE, i.e.

$$J^*_\beta(x) = \min_{u \in \mathbb{U}} \left[c(x, u) + \beta \sum_{y \in \mathbb{X}} J^*_\beta(y) T(y|x, u) \right]$$

- We define a Q-function as $Q_t : \mathbb{X} \times \mathbb{U} \to \mathbb{R}$ and the optimal Q-function, Q^*, as

$$Q^*(x, u) = c(x, u) + \beta \sum_{y \in \mathbb{X}} \min_v Q^*(y, v) T(y|x, u)$$

- Could iterate from Q_0 to get Q^*, but as mentioned earlier, this is difficult. So use Q-learning.
- For standard Q-learning, we fix an arbitrary policy, collect realizations of $X_t, U_t, c(X_t, U_t)$ and update our Q-functions as follows:

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

- Under some conditions on α_t, we get that this converges almost surely to the optimal Q function.

1.6 Quantized Q-learning

- For now, assume action space still finite but state space infinite. We wish to run Q-learning on some quantized version of this MDP.
- **PROBLEM:** We don't know if we run Q-learning on this finite state MDP that it will converge, and furthermore if it does converge we don't know if it converges to something meaningful (i.e. related to the original MDP).
- We construct a finite state approximation of the MDP as follows:
- Let $\{B_i\}_{i=1}^M$ be a partition of $\mathbb{X}$, and define a new (finite) state space as $\mathbb{Y}$, where $q(x) = y_i$ if $x \in B_i$.
- Define a "normalized weight measure" on B_i as

$$\hat{\pi}^*_{y_i}(A) := \frac{\pi^*(A)}{\pi^*(B_i)}$$

  for some $\pi^* \in \mathcal{P}(\mathbb{X})$ s.t. $\pi^*(B_i) > 0 \,\forall\, B_i$
- From this measure, we obtain a cost function and transition kernel for our finite-state model:

$$C^*(y_i,u) = \int_{B_i} c(x,u)\hat{\pi}^*_{y_i}(dx)$$

$$P^*(y_j|y_i,u) = \int_{B_i} T(B_j|x,u)\hat{\pi}^*_{y_i}(dx)$$

- In a similar fasion to above, we can define the DCOE for this model. Note that we can extend this function to $\mathbb{X}$.
- We also define $L : \mathbb{X} \to \mathbb{R}$ as

$$L(x) := \int_{B_i} ||x - x'||\hat{\pi}^*_{y_i}(dx)$$

  and

$$||L^-||_\infty := \max_i \sup_{x,x'\in B_i} ||x - x'||$$

- Note that we can find a partition s.t. this goes to 0 as $M \to \infty$.

Theorem 1.5. *If $T(\cdot|x,u)$ is weakly continuous in (x,u), and c is continuous and bounded, then we have for all compact $K \subset \mathbb{X}$*

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

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 the finite state MDP extended to $\mathbb{X}$

- This answers the second issue (i.e. if it converges, does it converge to something meaningful?). We are left with the question of whether it converges.
- We are considering the following algorithm

$$\begin{aligned} Q_{t+1}(q(x),u) = (1-&\alpha_t(q(x),u))Q_t(q(x),u)+ \\ &\alpha_t(q(x),u)(c(x,u) + \beta \min_{v \in \mathbb{U}} Q_t(q(X_{t+1}),v)) \end{aligned} \tag{3}$$

Assumption 1.1. *The following hold:*

1. $\alpha_t(y,u) = \frac{1}{\sum_{k=0}^{t} 1_{\{Y_k=y, U_k=u\}}}$ *if $(Y_t,U_t) = (y,u)$, and 0 otherwise.*
2. *Under a random exploration policy γ^*, $\{X_t\}_{t\geq 0}$ admits a unique invariant measure.*
3. *During exploration, every (y,u) is visited infinitely often.*

Theorem 1.6. *Under this assumption,* (3) *converges a.s. to*

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

where we get the C^* and P^* as above, where π^* is the invariant measure of $\{X_t\}_{t\geq 0}$ under γ^*.

1.7 Back to the Zero-Delay Coding Problem

- **IDEA:** Use quantized Q-learning, viewing π_t as the state and Q_t as the action.
- We want to see if the assumptions hold in the zero-delay coding problem. Indeed, we have weak continuity already, so we just need that under a random exploration policy, the "state" process $\{\pi_t\}_{t\geq 0}$ admits a unique invariant measure.