Reinforcement Learning for Jointly Optimal Coding and Control over a Communication Channel

Evelyn Hubbard Liam Cregg Serdar Yüksel

Queen's University

Problem Set-Up

A Markov source $(x_t)_{t\geq 1} \in \mathcal{X}$, holds the state process and is updated at a plant with dynamics:

$$x_{t+1} = f(x_t, u_t, w_t)$$

- $u_t \in \mathcal{U}$ is a control action and w_t is i.i.d noise.
- x_t is causally encoded (quantized) to a quantization output $q_t \in \mathcal{M}$.
- q_t is fed over a noiseles finite-rate channel to the controller, which chooses an action u_t.

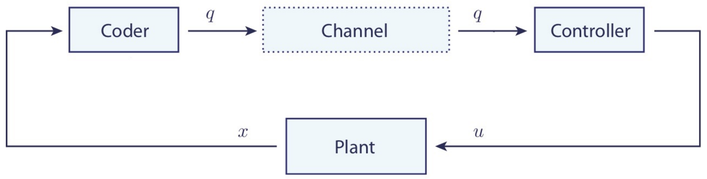


Figure 1. Control driven Markov process over Noiseless Channel

Policies are sequences of functions, which generate q_t and u_t.

- An admissible *coding policy* is $\gamma^e = (\gamma_t^e)_{t\geq 1}$, where for each t:
$$\gamma_t^e : \{x_{[0,t]}, q_{[0,t-1]}\} \mapsto \mathcal{M}.$$
- An admissible *control policy* is $\gamma^c = (\gamma_t^c)_{t\geq 1}$, where for each t:
$$\gamma_t^c : \{q_{[0,t]}, u_{[0,t-1]}\} \mapsto \mathcal{U}.$$
- A joint policy is $\overline{\gamma} = \{\gamma^e, \gamma^c\}$.

Optimization Objective (for an initial distribution π_0):

$$J^N(\pi_0) = \inf_{\overline{\gamma}\in\Gamma_A} J^N(\pi_0, \overline{\gamma}) := \inf_{\overline{\gamma}\in\Gamma_A} E_{\pi_0}^{\overline{\gamma}}\left[\sum_{k=0}^{N-1} c(x_k, u_k)\right]$$

where Γ_A is the set of joint admissible policies.

Controlled Predictor Structured Policies

Predictor sequence on $\mathcal{X}$: $\pi_t(\cdot) := P^{\overline{\gamma}}(x_t = \cdot\,|q_{[0,t-1]})$
We extend Walrand and Variya's results to define a new policy class [1, 2, 3, 4].

- any admissible coding policy can be replaced without loss in performance by one which uses only π_t and x_t
$$Q_t = \gamma_t^e(\pi_t),\ q_t = Q_t(x_t)$$
where $Q_t \in \mathcal{Q}$ represents a quantizer.
- any control policy can be replaced without loss in performance by one which uses only π_t, q_t, and Q_t.
$$\eta_t = \gamma_t^c(\pi_t),\ u_t = \eta_t(q_t, Q_t)$$
where $\eta_t \in \mathcal{H} = \{\mathcal{Q}\times\mathcal{M} \mapsto \mathcal{U}\}$ represents a map to control actions.

A policy, $\overline{\gamma}$, that uses information in this way has *Controlled-Predictor Structure* ($\overline{\gamma} \in \Gamma_{C-P}$).

Structural Results: Markov Decision Process Reformulation

Theorem 1: $(\pi_t, (Q_t, \eta_t))$ is a controlled Markov chain with π_t as the effective state and (Q_t, η_t) as the effective control action.

Theorem 2: The transition kernel $\tilde{P}(\pi_{t+1}|\pi_t, Q_t, \eta_t)$ is weakly continuous in $\mathcal{P}(\mathcal{X})\times\mathcal{Q}\times\mathcal{H}$.

The optimization objective for the new MDP is

$$J^N(\pi_0) := \inf_{\overline{\gamma}\in\Gamma_{C-P}} E_{\pi_0}^{\overline{\gamma}}\left[\sum_{k=0}^{N-1} \tilde{c}(\pi_k, Q_k, \eta_k)\right].$$

Theorem 3: The minimum cost of the original and reformulated problems are equivalent.

$$\inf_{\overline{\gamma}\in\Gamma_A} E_{\pi_0}^{\overline{\gamma}}\left[\sum_{k=0}^{N-1} c(x_k, u_k)\right] = \inf_{\overline{\gamma}\in\Gamma_{C-P}} E_{\pi_0}^{\overline{\gamma}}\left[\sum_{k=0}^{N-1} \tilde{c}(\pi_t, Q_t, \eta_t)\right].$$

The minimum infinite horizon discounted cost criteria is:

$$J_\beta(\pi_0) = \inf_{\overline{\gamma}\in\Gamma_A} J_\beta(\pi_0, \overline{\gamma}) = \inf_{\overline{\gamma}\in\Gamma_A} \lim_{N\to\infty} E^{\overline{\gamma}}\left[\sum_{k=0}^{N-1} \beta^k \tilde{c}(\pi_k, Q_k, \eta_k)\right] \quad (1)$$

where,

$$\tilde{c}(\pi_t, Q_t, \eta_t) = \sum_{\mathcal{X}\times\mathcal{U}\times\mathcal{M}} \eta_t(u_t|q_t, Q_t) Q_t(q_t|x_t)\pi_k(x_t)c(x_t, u_t).$$

Theorem 4: An optimal stationary policy for (1) exists in Γ_{C-P}.

Sliding Finite Window Structure Markov Problem and Approximation

Instead of π_t, consider $\kappa_t = \{\pi_{t-N}, I_t\}$ to be the state, where:

$$I_t^N = \{q_{[t-N,t-1]}, Q_{[t-N,t-1]}, \eta_{[t-N,t-1]}\}.$$

- A joint coding-controller policy has *Controlled Finite Sliding Window Structure* (Γ_{C-FW}) if, at time t, it uses only w_t to select Q_t and η_t.
- An approximation for w_t, which starts from an incorrect prior μ is:
$$\hat{\kappa}_t = (\mu, I_t^N).$$
- This approximation is a good one if some filter stability conditions are met (ensuring the process can recover from starting at the incorrect prior).
- $(\hat{\kappa}, Q, \eta)$ is a **Window-Length-N Approximation** for a finite sliding window MDP, (MDP_N).

Theorem 5: If the optimal policy for the MDP_N, γ_N is extended over $\mathcal{P}(\mathcal{X})$. Then for any $\gamma \in \Gamma_{C-P}$ which is applied N times starting from π_{-N} to generate κ_0, we have

$$E_{\pi_{t-N}}^{\gamma}[\dot{J}_\beta(\kappa_0, \gamma_N) - \dot{J}_\beta(\kappa_0)] \leq \frac{4||c||_\infty}{(1-\beta)^2}(2(1 - \min_{u\in\mathcal{U}} \delta(P, u)))^N$$

and if $\min_{u\in\mathcal{U}} \delta(P, u) > 1/2$,

$$\lim_{N\to\infty} |\dot{J}_\beta(\kappa_0, \gamma_N) - \dot{J}_\beta(\kappa_0)| = 0$$

where:

- $\dot{J}_\beta(\kappa_0) = \inf_{\gamma\in\Gamma_{C-FW}} \dot{J}_\beta(\kappa_0, \gamma)$ is a minimum cost for the finite sliding window structure
- $||c||_\infty := \max_{(x,u)\in\mathcal{X}\times\mathcal{U}} c(x, u)$
- $\delta(P, u)$ is the Dobrushin coefficient of $P(\cdot|\cdot, u)$

Finite State Approximation via Predictor Quantization

Quantized predictors, $\hat{\pi}$, exist in a finite predictor space:

$$P_n(\mathcal{X}) = \left\{\hat{\pi} \in P(\mathcal{X}) \colon \hat{\pi} = \left[\frac{k_1}{n}, \ldots, \frac{k_{|\mathcal{X}|}}{n}\right], k_i = 0, \ldots, n, i = 1, \ldots, |\mathcal{X}|\right\}.$$

- $(\hat{\pi}, Q, \eta)$ on $P_n(\mathcal{X})\times\mathcal{Q}\times\mathcal{H}$ is an **n-Level Finite State Approximation** for an MDP, defined MDP_n.
- The cost function for the MDP_n, $\tilde{c}_n(\hat{\pi}, Q, \eta)$ is bounded and the kernel, $\tilde{P}_n(\hat{\pi}_j|\hat{\pi}_i, Q, \eta)$ is weakly continuous.

Theorem 6 [5]: **If the optimal policy for the MDP_n, γ_n is extended over the entire MDP (by remaining constant in the quantizer cells), it becomes near-optimal for the original MDP.**

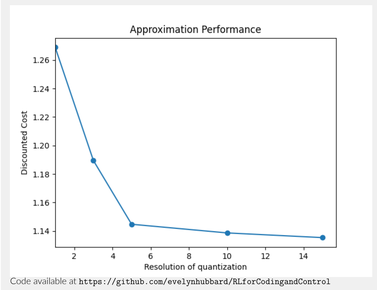
$$\lim_{n\to\infty} \left|J_\beta(\pi_0, \gamma_n) - J_\beta(\pi_0)\right| = 0$$

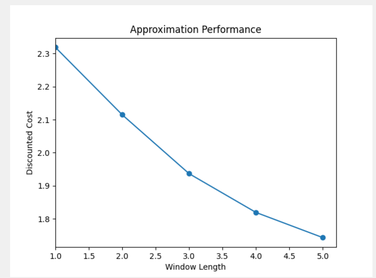
Q-Learning Algorithm

- Theorems 5 and 6 imply that solutions found by Q-Learning algorithms on finite-state-approximate-MDPs will be near-optimal for the original MDPs.
- Q-learning iterations for every state-action pair (say $(y, u) \in \mathcal{Y}\times\mathcal{U}$) would be:
$$\mathbf{Q}_{t+1}(y, u) = (1 - \alpha_t(y, u))\mathbf{Q}_t(y, u) + \alpha_t(y, u)(c_t + \beta \min_{u\in\mathcal{U}} \mathbf{Q}(y_{t+1}, u))$$
- Both MDP approximations meet the assumptions required for the following convergence theorem for Q-learning in non-Markovian environments, from [5].

Theorem 7:

1. $\mathbf{Q}_t(y, u) \to \mathbf{Q}^*(y, u)$ almost surely for each $(y, u) \in \mathcal{Y}\times\mathcal{U}$.
2. $\mathbf{Q}^*(y, u)$ is the solution to
$$\mathbf{Q}^*(y, u) = c(y, u) + \beta \sum_{y'\in\mathcal{Y}} \min_{u'\in\mathcal{U}} \mathbf{Q}(y', u')P(y'|y, u)$$
3. An optimal policy for $MDP_n = (\mathbb{Y}, \mathbb{U}, P, c)$ is given by: $\gamma_n(y) := arg\min_{u\in\mathcal{U}} \mathbf{Q}^*(y, u)$.





Code available at `https://github.com/evelynhubbard/RLforCodingandControl`

References

[1] J. C. Walrand and P. Varaiya. Optimal causal coding-decoding problems. *IEEE Transactions on Information Theory*, 19(11):814–820, 1983.

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

[3] A. Mahajan and D. Teneketzis. Optimal performance of networked control systems with non-classical information structures. *SIAM Journal of Control and Optimization*, 48:1377–1404, May 2009.

[4] R.G. Wood, T. Linder, and S. Yüksel. Optimal zero delay coding of Markov sources: Stationary and finite memory codes. *IEEE Transactions on Information Theory*, 63:5968–5980, 2017.

[5] A.D 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*, pages 1–34, 2023.

[6] A. D. Kara and S. Yüksel. Q-Learning for Stochastic Control under General Information Structures and Non-Markovian Environments. *Transactions on Machine Learning Research (arXiv:2311.00123)*, 2024.

[7] L. Cregg, T. Linder, and S. Yüksel. Reinforcement learning for near-optimal design of zero-delay codes for markov sources. *IEEE Transactions on Information Theory, arXiv:2311.12609*, 2024.