Q-LEARNING FOR NON-MARKOVIAN ENVIRONMENT

VIVEK S. BORKAR[*]

IIT BOMBAY

EECS@IISc, April 3, 2023

[*] Joint work with Siddharth Chandak and Parth Dodhia, IIT Bombay (now at Stanford Uni.), Pratik Shah (still at IITB)

The Van Roy formalism for RL in arbitrary environments (Dong, Van Roy, Zhou, 2021)

Basic components:

Observed process taking values in a space $\mathcal{O}$, to be controlled.

Agent state $\{S_n\}$ taking values in a state space $\mathcal{S}$.

Control sequence $\{U_n\}$ taking values in an action space $\mathcal{U}$.

The state is given recursively by

$$S_{n+1} = f(S_n, U_n, O_{n+1}), \; n \geq 0,$$

for some $f : \mathcal{S} \times \mathcal{U} \times \mathcal{O} \mapsto \mathcal{S}$.

At time n, a reward $r(S_n, U_n, O_{n+1})$ is received.

Main result: a regret bound for the reward.

Refined model in (Lu, Van Roy, Dwaracherla, Ibrahimi, Osband, Wen, 2021)

Our objective:

We take a somewhat different take on this problem. In order to highlight the main differences, we work with the simpler set up of (Dong et al.), though the results can be easily extended to the more general framework of (Lu et al.).

In addition, we take the observation process O_n to have been evolving forever, i.e., for $n > -\infty$.

$O^n :=$ the semi-infinite sequence $[O_n, O_{n-1}, \cdots] \in \mathcal{O}^\infty$

$U^n :=$ the semi-infinite sequence $[U_n, U_{n-1}, \cdots] \in \mathcal{U}^\infty$

$H_n := (O^n, U^n), n \geq 0.$

We drop the explicit randomization ξ_{n+1} and subsume it in a possibly randomised control U_n.

$\mathcal{P}(\cdots) :=$ the space of probability measures on the Polish space '$\cdots$'.

We take $\mathcal{S}, \mathcal{U}, \mathcal{O}$ to be finite sets.

Our main contributions:

1. By explicitly pinning down the error caused by potential non-Markovianity of the observation process in the application of the Q-learning algorithm, we propose an alternative performance criterion.

2. We argue that this reduces the problem to seeking good approximations for certain conditional laws.

3. We justify achieving such approximation by ‘recursively computable approximate sufficient statistics’, drawing upon the notion of Partially Observed Markov Decision Processes (POMDPs) in classical stochastic control.

4. We propose a scheme for this purpose inspired by autoencoders.

At a meta level, our attempt is to make a connection between reinforcement learning and classical stochastic control.

In the process, we also highlight a parallel with the classical 'internal model principle' of control theory that deserves to be better appreciated as a guiding principle.

THE CURSE OF NON-MARKOVIANITY

Our agent dynamics is given by

$$S_{n+1} = f(S_n, U_n, O_{n+1}), \; n > -\infty.$$

We assume that $\{(S_n, U_n, O_n)\}$ is stationary ('asymptotically stationary' will do) and satisfies

$$P(O_{n+1} \in A | O^n, U^n) = p(A | O^n, U^n)$$

for a continuous map

$$(z, a) \in \mathcal{O}^\infty \times \mathcal{U}^\infty \mapsto p(\cdot | z, a) \in \mathcal{P}(\mathcal{O}).$$

We assume that the state space $\mathcal{S}$ and the action space $\mathcal{U}$ are finite with cardinalities $\mathfrak{s}$ and $\mathfrak{u}$, respectively.

Let $r(O^{n+1}, U^n)$ be the reward at time n and

$$\bar{r}(i, u) := E[r(O^{n+1}, U^n)|S_n = i, U_n = u].$$

Furthermore, assume '**stationary randomized policy**': the conditional law of U_n given $S_m, O_m, U_{m-1}, m \leq n$, is some $\phi(\cdot|S_n)$.

Consider the Q-learning algorithm applied to the agent dynamics above, disregarding the non-Markovianity of $\{S_n\}$.

This leads to:

$$Q_{n+1}(i,u) = Q_n(i,u) + a(n)I\{S_n = i, U_n = u\}\times$$
$$\Big[r(O^{n+1},U^n) + \gamma \max_a Q_n(S_{n+1},a) - \ Q_n(i,u)\Big].$$

Let $q(s'|s,u) := P(S_{n+1} = s'|S_n = i, U_n = u)$, where the absence of explicit time dependence of q is due to our stationarity hypothesis. Let $\pi(O^n,U^n)$ and $\tilde{\pi}(S_n,U_n)$ denote the stationary distributions of (O^n,U^n) and (S_n,U_n), respectively.

Important observation: The stationary law φ of (S_n,U_n) is also the stationary law of the $\mathcal{S}\times\mathcal{U}$-valued Markov chain with transition probability $\psi(i',u'|i,u) := q(i'|i,u)\phi(u'|i)$.

In particular, the corresponding parametrized Poisson equation holds:

$$\begin{aligned} V(x,i,u) &= F(x,i,u) - \sum_{i',u'} \tilde{\pi}(i',u') F(x,i',u') + \\ &\quad \sum_{i',u'} \psi(i',u'|i,u) V(x,i',u'), \end{aligned}$$

where,

$$\begin{aligned} F^{s,v}(x,i,u) &:= I\{i=s, u=v\}\Big(\bar{r}(s,v) + \\ &\quad \gamma \sum_{i',v'} q(i'|s,v) \max_w x(i',w) - x(s,v)\Big), \end{aligned}$$

is the $(s,v)^{\text{th}}$ element of $F(x,i,u)$.

Using the Poisson equation, the scheme can be written as

$$\begin{aligned} Q_{n+1}(i,u) \;=\; & Q_n(i,u) + a(n) I\{S_n = i, U_n = u\} \times \\ & \Big[\bar{r}(i,u) \;+ \gamma \sum_j q(j|i,u) \max_a Q_n(j,a) \\ & - Q_n(i,u) + \; M_{n+1} + \; \zeta_{n+1} + \; \varepsilon_{n+1}\Big]. \end{aligned}$$

Here,

1. ε_n is an asymptotically vanishing random bounded sequence arising from the Poisson equation,

2.

$$\begin{aligned} M_{n+1} &:= \gamma(\max_a Q_n(S_{n+1}, a) \\ &\quad - \sum_j \max_a Q(j, a) P(S_{n+1} = j | O^n, U^n)) \\ &\quad + \, (r(O^{n+1}, U^n) - E[r(O^{n+1}, U^n) | O^n, U^n]) \\ &\quad + \, (V(Q_n, S_{n+1}, U_{n+1}) - E\left[V(Q_n, S_{n+1}, U_{n+1}) | O^n, U^n\right]) \end{aligned}$$

is a martingale difference sequence, and, 3.

$$\begin{aligned}\zeta_{n+1} := \gamma(&\sum_j \max_a Q(j,a) P(S_{n+1} = j | O^n, U^n) \\ &- \sum_j q(j|S_n, U_n) \max_a Q_n(j,a) \\ &+ \ (E[r(O^{n+1}, U^n)|O^n, U^n] - \bar{r}(S_n, U_n)) \\ &- \ (E\left[V(Q_n, S_{n+1}, U_{n+1})|S_n, U_n\right] \\ &- \ E\left[V(Q_n, S_{n+1}, U_{n+1})|O^n, U^n\right])\end{aligned}$$

is the offset due to non-Markovianity.

The term ζ_n is due to non-Markovianity and is not present in the classical analysis of Q-learning.

For $n \geq 0$, define:

$$\begin{aligned} b_k(n) &= \sum_{m=k}^{n} a(m), \ 0 \leq k \leq n < \infty, \\ \beta_k(n) &= \begin{cases} \frac{1}{k^{d_2-d_1} n^{d_1}}, & \text{if } d_1 \leq d_2 \\ \frac{1}{n^{d_2}}, & \text{otherwise,} \end{cases} \\ \kappa(d) &= \|\mathbf{1}\|, \ \mathbf{1} := [1, 1, \cdots, 1]^T \in \mathcal{R}^d, \ d \geq 1, \\ \Delta(n) &= \sum_{m=n_0}^{n} \prod_{k=m+1}^{n} (1 - a(k)) a(m+1) \zeta_{m+1}. \end{aligned}$$

Let Q^* denote the unique solution to the equation

$$Q^*(i,u) = \bar{r}(i,u) + \gamma \sum_j q(j|i,u) \max_v Q^*(j,v).$$

Theorem 1 $Q_n \to Q^*$ a.s. Furthermore, let $n_0 \geq 0$ satisfy $a(n_0) < 1$, $a(n)$ is non-increasing after n_0 and $\frac{d_1}{n} \leq a(n) \leq d_3 \left(\frac{1}{n}\right)^{d_2}, \forall\ n \geq n_0$. Then there exist finite positive constants c_1, c_2 and D, depending on $\|Q_{n_0}\|$, such that for $\delta > 0$ and $n \geq n_0$, the inequality

$$\begin{aligned}\|Q_n - Q^*\| \ \leq \ & e^{-(1-\alpha)b_{n_0}(n)}\|Q_{n_0} - Q^*\| \\ & + \frac{\delta + a(n_0)c_1}{1-\alpha} + \Delta(n)\end{aligned}$$

holds with probability exceeding

$$\begin{array}{ll} 1 \ - \ 2\mathfrak{su} \displaystyle\sum_{m=n_0+1}^{n} e^{-D\delta^2/\beta_{n_0}(m)}, & 0 < \delta \leq C, \\ 1 \ - \ 2\mathfrak{su} \displaystyle\sum_{m=n_0+1}^{n} e^{-D\delta/\beta_{n_0}(m)}, & \delta > C. \end{array}$$

Here $C = e^{\left(2\left(1+\|Q_{n_0}\|_\infty+\frac{1}{1-\alpha}\right)+c_2\right)}$ and $\alpha = 1-(1-\gamma)\pi_{min}$, where $\pi_{min} = \min_{i,u}\tilde{\pi}(i,u)$.

This follows exactly as in Corollary 2 of (Chandak, Borkar, Dodhia, *Stochastic Systems*, 2021). The only difference is the additional error term $\Delta(n)$ due to non-Markovianity which was not present in the bound of *ibid.*, because the latter considered only the Markovian case.

From Chandak et al., we also have the following.

- The first term on the right in the above bound decays to zero as $n \uparrow \infty$ because of our assumptions on $\{a(n)\}$.

- Results of Chandak et al. also show that the second term on the right can be made arbitrarily small by setting $\delta = \delta(n_0) \downarrow 0$ as $n_0 \uparrow \infty$.

Thus it is only the error due to non-Markovianity, $\Delta(n)$, that will matter as $n \uparrow \infty$. We show that $\Delta_n \to 0$ a.s., but a convenient concentration bound seems unavailable except possibly in very special cases. Even when available, it is expected to be rather weak.

Our objective is to use this to motivate a particular performance criterion for agent design.

We next introduce this performance criterion and the ensuing notion of *Recursively Computable Approximate Sufficient Statistics*.

Recursively Computable Approximate Sufficient Statistics

In view of the above, a legitimate criterion for agent design is to target the error $\{\Delta(n)\}$ and our objective should be to minimize it.

This implies that we need the conditional distribution of (S_{n+1}, O_{n+1}) given (O^n, U^n) to be well approximated by its conditional distribution given (S_n, U_n). In other words, we need S_n to be an ‘approximate sufficient statistics’.

It is instructive here to draw a parallel with the theory of POMDPs in classical stochastic control.

To make the parallels apparent, we shall repeat the same notation.

In a POMDP, one controls a (controlled) Markov chain $\{S_n\}$ on a finite state space $\mathcal{S}_3$ given a process of observations $\{O_n\}$ taking values in a finite space, with controlled transition kernel (say) $p'(j, o|i, u), i, j \in \mathcal{S}$ satisfying $\forall n$,

$$P(S_{n+1} = j, O_{n+1} = o|S^n, U^n, O^n) = p'(j, o|S_n, U_n).$$

The control U_n is allowed to depend only upon O^n, U^{n-1}, and independent extraneous randomization.

Then $\pi_n(\cdot) :=$ the conditional law of S_n given (O^n, U^n), is given recursively by the nonlinear filter

$$\pi_{n+1}(i) = \frac{\sum_j \pi_n(j)p'(i, O_{n+1}|j, U_n)}{\sum_{j,k} \pi_n(j)p'(k, O_{n+1}|j, U_n)}, \ \forall n.$$

This is an easy consequence of the Bayes rule.

Then $\{\pi_n\}$ is a $\mathcal{P}(\mathcal{S})$-valued ‘completely observed’ controlled Markov chain (separated control problem).

The above recursion has the same form as our dynamics for $\{S_n\}$, except that π_n resides in the probability simplex $\mathcal{P}(\mathcal{S})$, which is not a finite set. Thus replacing it with the latter dynamics amounts to a finite state approximation to $\{\pi_n\}$.

This is the point of view taken in, e.g., Yu and Bertsekas (2008), where near-optimality of finite state controllers is explored.

A critical difference in our case is that $\{\pi_n\}$ is *not* a controlled Markov chain.

This is the crux of the *curse of non-Markovianity*.

It is worth noting that under appropriate conditions on the controlled Markov chain used for sampling, the classical Q-learning algorithm will converge regardless, but not necessarily to the desired limit.

Considering the asymptotic stationary regime as $n \uparrow \infty$, consider the controlled Markov chain $(O^n, U^{n-1}), n > -\infty$, controlled by the control process $U_n, n > -\infty$.

Then the agent state S_n is perforce of the form $S_n = F(O^n, U^{n-1})$ for a measurable $F : \mathcal{O}^\infty \times \mathcal{U}^\infty \mapsto \mathcal{S}$.

Also, the sets $\{S_n = i, U_{n-1} = u\}$ for $i \in \mathcal{S}$ is a partition of $\mathcal{O}^\infty \times \mathcal{U}^\infty$ independent of n by stationarity.

Recall that in the transition $(O^n, U^{n-1}) \to (O^{n+1}, U^n)$, only the last component of the two strings gets replaced.

In fact, U_n is generated according to a fixed conditional law $\tilde{\varphi}(\cdot|O^n, U^{n-1})$ given (O^n, U^{n-1}), conditionally independent of all else.

That is, we are implementing a 'stationary randomized policy' for the chain (O^n, U^{n-1}).

Analogy with Singh, Jaakkola, Jordan (1994) (see also Tsitsiklis and Van Roy, 1996) suggests that the following should hold:

Under our hypotheses, $\{Q_n\}$ converge a.s. to the unique solution of the system of equations

$$Q(s,u) = \sum_{z \in \mathcal{O}^\infty \times \mathcal{U}^\infty} P((O^0, U^0) = z | s,u) \times \\ \left(\sum_{o' \in \mathcal{O}, s' \in \mathcal{S}} p(s', o' | z) \Big(r(s,u,o') + \gamma \max_a Q(s', a) \Big) \right).$$

This turns out to be equivalent to our convergence result.

Note that $\{O_n\}$ is trivially a POMDP if we view $X_n :=$ (O^n, U^{n-1}) as a controlled Markov chain with state space $\mathcal{S}_2 := \mathcal{O}^\infty \times \mathcal{U}^\infty$, with the associated observation process $O_n = g(X_n)$ where

$$g : ((o_n, u_{n-1}), (o_{n-1}, u_{n-2}), \cdots) \in (\mathcal{O} \times \mathcal{U})^\infty \mapsto o_n \in \mathcal{O}.$$

Thus the agent design task can also be viewed as developing an approximate *finite state* model thereof.

More generally, it is the problem of developing and then controlling a surrogate simpler model of the original system. This is reminiscent of the *internal model principle* in adaptive control, an informal statement of which says:

‘a good controller incorporates a model of the dynamics that generate the signals which the control system is intended to track’.

A natural choice for this purpose that comes to mind is approximation of (O^n, U^{n-1}) by a finite window

$$[O_{n-K}, U_{n-K}, \ldots, O_{n-1}, U_{n-1}, O_n].$$

This is as a special case of the framework described above, extensively analyzed by (Kara and Yuksel, 2021), who propose a Q-learning algorithm in this framework.

Another alternative approach to learning RCASS by Subramanian, Sinha, Seraj, Mahajan (2021) borrows from the reinforcement learning paradigm.

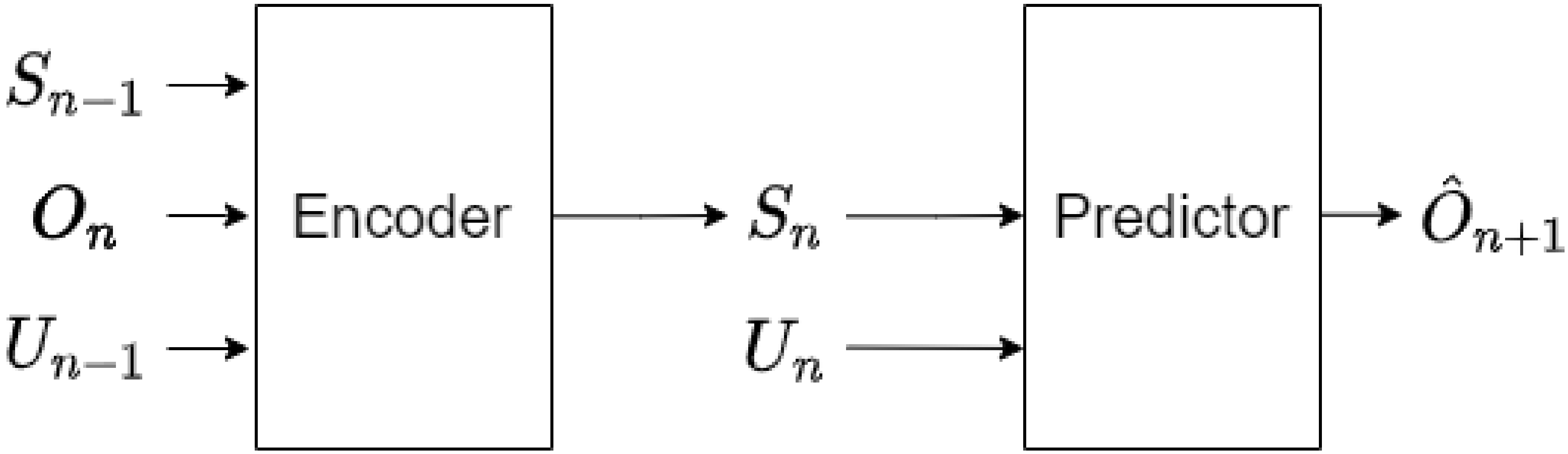
An ‘autoencoder’

We now propose a specific architecture inspired by Champion, Lusch, Kutz, Brunton (2019).

However, unlike the *auto*-encoder of *ibid.* where one tries to match the output with the input, we match the output with the *next* input, i.e., our error is not to be viewed as a decoding error, but as a prediction error.

Our architecture has a ‘state’ neural network that takes S_n, U_n, O_{n+1} as input and gives the next agent state S_{n+1} as output.

This feeds into the ‘observation’ neural network that takes this S_{n+1} along with U_n as input and outputs a prediction $\overline{O}_{n+1}$ of the next observation O_{n+1} as the output.


S_{n-1}
O_n
U_{n-1}
Encoder
S_n
U_n
Predictor
$\hat{O}_{n+1}$

The neural networks are trained by comparing $O_{n+1}, \overline{O}_{n+1}$. The criterion for this comparison can be chosen depending on the problem specifics.

For example, it could be one of the classical loss functions penalizing the difference between O_{n+1} and $\overline{O}_{n+1}$, or it may compare their distributions or conditional distributions.

A few simple numerical experiments using least squares error criterion give promising results.

Example - Cartpole

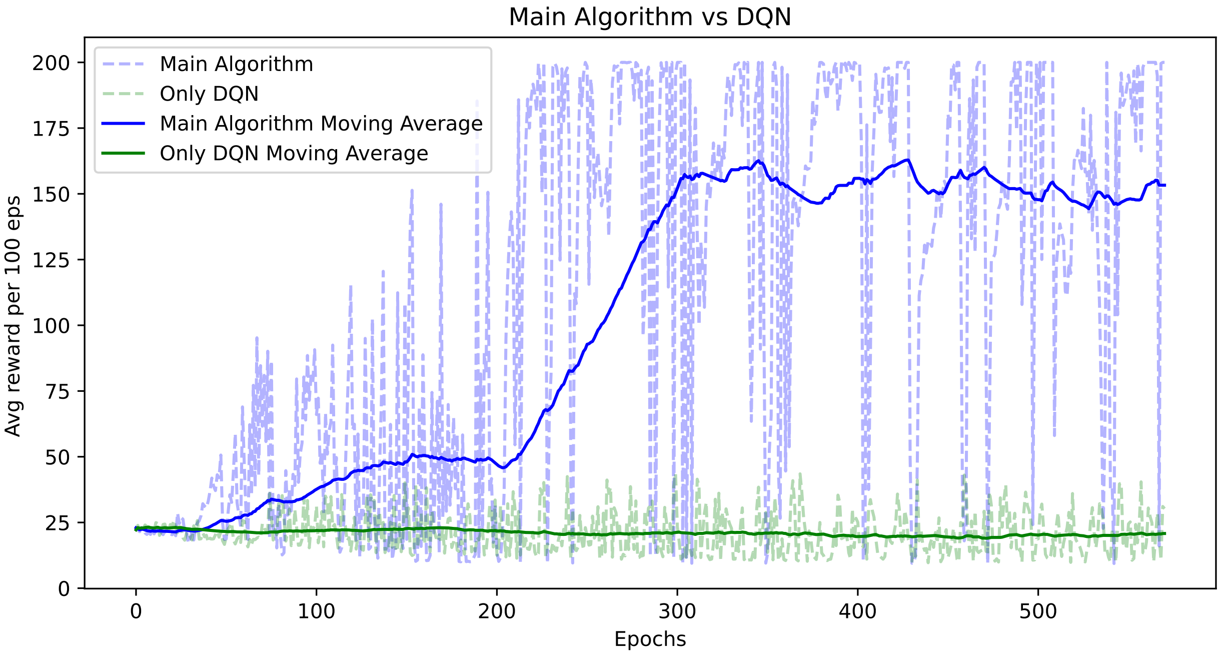
The environment consists of a cart with an upside-down pole where the agent has to maintain the vertical angle (α_n) of the pole sufficiently low while remaining in a restricted horizontal region (x_n).

System state is $X_n = (\alpha_n, \omega_n, x_n, v_n)$ where α_n and x_n are as above and ω_n is the angular velocity and v_n is the horizontal velocity.

The observation process is $O_n = (\alpha_n, x_n)$.

The possible actions are to apply force either to the left or to the right.

We consider an episodic environment and take $\gamma = 1$. The agent gets a reward of 1 until the episode ends or the constraints are violated. The maximum reward agent can attain in an episode is 200.


Main Algorithm vs DQN
Main Algorithm
Only DQN
Main Algorithm Moving Average
Only DQN Moving Average
Avg reward per 100 eps
0
25
50
75
100
125
150
175
200
Epochs
0
100
200
300
400
500