Q-LEARNING WITHOUT STOCHASTIC APPROXIMATION

Vivek S. Borkar, IIT Bombay*†

Mar. 23, 2015, IIT, Chennai

*Joint work with Dileep Kalathil (Uni. of California, Berkeley), Rahul Jain (Uni. of Southern California)

†Work supported in part by the Department of Science and Technology

OUTLINE

1. Markov Decision Processes (Discounted cost)

2. Value/Q-value iteration algorithms

3. Classical Q-learning

4. Main results

$\{X_n, n \geq 0\}$ a *controlled* Markov chain with:

- a finite state space $S = \{1, 2, \cdots, s\}$,
- a finite action space $A = \{a_1, \cdots, a_d\}$,
- an A-valued control process $\{Z_n, n \geq 0\}$,

- a controlled transition probability function

$$p(j|i,u),\ \ i,j \in S, u \in A,$$

such that

$$P(X_{n+1} = i|X_m, Z_m, m \leq n) = p(i|X_n, Z_n)\ \forall n,$$

i.e., the probability of going from $X_n = j$ (say) to i under action $Z_n = u$ (say) is $p(i|j,u)$.

Say that $\{Z_n\}$ is:

- *admissible* if above holds,

- *randomized stationary Markov* if

$$P(Z_n = u|\mathcal{F}_{n-1}, X_n = x) = (\varphi(x))(u) \ \forall n$$

  for some $\varphi : S \mapsto \mathcal{P}(A)$,

- *stationary Markov* if $Z_n = v(X_n) \ \forall n$ for some $v : S \mapsto A$.

With abuse of terminology, the last two are identified with φ, v esp.

Objective: Minimize the discounted cost

$$J_i(\{Z_n\}) := E\left[\sum_{m=0}^{\infty} \beta^m c(X_m, Z_m) | X_0 = i\right],$$

where

- $c : S \times A \mapsto \mathcal{R}$ is a prescribed 'running cost' function,

- $\beta \in (0, 1)$ is the discount factor.

Dynamic Programming

Define '*value function*' $V : S \mapsto \mathcal{R}$ by

$$V(i) = \inf_{\{Z_n\}} J_i(\{Z_n\}).$$

Then by the 'dynamic programming principle'

$$V(i) = \min_u \left[c(i,u) + \beta \sum_j p(j|i,u) V(j) \right], \ i \in S.$$

This is the associated *dynamic programming equation*.

Furthermore, if the minimum of the right is attained at $u = v^*(i)$, then the stationary Markov policy $v^*(\cdot)$ is optimal. The converse also holds.

DP equation is a fixed point equation: $V = F(V)$ for

$$\begin{aligned} F(x) &= [F_1(x), \cdots, F_s(x)]^T \text{ where} \\ F_i(x) &:= \min_u [c(i,u) + \beta \sum_j p(j|i,u) x_j]. \end{aligned}$$

Then $\|F(x) - F(y)\|_\infty \leq \beta \|x - y\|_\infty$, i.e., F is an $\|\cdot\|_\infty$-contraction

$\Longrightarrow V$ a unique solution to the DP equation and the '*value iteration scheme*'

$$V^{n+1}(i) = \min_u \left[c(i,u) + \beta \sum_j p(j|i,u) V^n(j) \right], \ n \geq 0,$$

converges exponentially to V.

Other schemes: policy iteration, linear programming (primal/dual)

Problematic if:

- (i) $p(\cdot|\cdot,\cdot)$ unknown, or,

- $(ii) p(\cdot|\cdot,\cdot)$ known, but too complex (e.g., extremely large state space).

Sometimes simulation of the system is 'easy', e.g., when the system is composed of a large number of interconnected simple components whose individual transitions are easy to simulate

(e.g., queuing networks, robots).

This has motivated simulation based schemes for approximate dynamic programming, based on stochastic approximation versions of classical iterative schemes.

('reinforcement learning', 'approximate dynamic programming', 'neurodynamic programming')

Q-learning: a simulation based scheme for approximate dynamic programming due to CJCH Watkins (1992).

Define *Q-values*

$$Q(i,u) := c(i,u) + \beta \sum_j p(j|i,u) V(j), \;\; i \in S, u \in A.$$

Then

$$\begin{aligned} V(i) &= \min_u Q(i,u), \\ Q(i,u) &= c(i,u) + \beta \sum_j p(j|i,u) \min_a Q(j,a). \end{aligned}$$

This is the 'DP equation' for Q-values.

Again, the last equation is of the form $Q = G(Q)$ where

$$\|G(x) - G(y)\|_\infty \leq \beta \|x - y\|_\infty$$

Thus we have the '*Q-value iteration*'

$$Q^{n+1}(i,u) = c(i,u) + \beta \sum_j p(j|i,u) \min_a Q^n(j,a), \ \ n \geq 0.$$

Then $Q^n \rightarrow$ the unique solution to the Q-DP equation.

Furthermore, $v^*(i) \in$ Argmin $Q(i,\cdot), i \in S$, yields an optimal stationary Markov policy v^*.

Note $V^n \in \mathcal{R}^s, \ Q^n \in \mathcal{R}^{s \times d} \Longrightarrow$ no motivation to do Q-value iteration.

However, one big change from value iteration:

the nonlinearity (minimization over A) is now inside the averaging

$\Longrightarrow$ can use an incremental method based on stochastic approximation.

Advantage: can be based upon simulation,
low computation per iterate

Disadvantage: slow convergence

Stochastic Approximation

Robbins-Monro scheme:

$$x(n+1) = x(n) + a(n)[h(x(n)) + M(n+1)].$$

Here, for $\mathcal{F}_n := \sigma(x(0), M(k), k \leq n)$ (i.e., the 'history till time n'),

- $a(n) > 0$ with $\sum_n a(n) = \infty,\ \sum_n a(n)^2 < \infty$, and,

- $\{M(n)\}$ a martingale difference sequence:

$$E[M(n+1)|\mathcal{F}_n] = 0\ \forall n.$$

Need: h Lipschitz and

$$E[\|M(n+1)\|^2|\mathcal{F}_n] \leq K(1+\|x(n)\|^2).$$

Typically,

$$x(n+1) = x(n) + a(n)f(x(n), \xi(n+1)),$$

with $\{\xi(n)\}$ IID. Then set

$$\begin{aligned} h(x) &= E[f(x, \xi_n)], \\ M(n+1) &= f(x(n), \xi(n+1)) - h(x(n)). \end{aligned}$$

'ODE' approach (Derevitskii-Fradkov, Ljung):

Treat the iteration as a noisy discretization of the ODE

$$\dot{x}(t) = h(x(t)).$$

If this has x^* as its unique asymptotically stable equilibrium, then

$$\sup_n \|x(n)\| < \infty \Longrightarrow x(n) \to x^* \text{ a.s.}$$

(LHS needs separate 'stability' tests)

Idea of proof:

Treat the iteration as noisy discretization of the ODE.

Specifically,

- define $\bar{x}(t), t \geq 0$, by $\bar{x}(\sum_{m=0}^{n-1} a(m)) := x(n)$, with linear interpolation,

- compare $\bar{x}(s), t \leq s \leq t + T$, with ODE trajectory on the same time interval with the same initial condition,

- Gronwall inequality yields bound in terms of discretization error and error due to noise,

- verify that these errors go to zero asymptotically (the latter follows by martingale arguments, using square-summability of $\{a(n)\}$),

- use either a Liapunov function argument (when available) or a characterization of limit set (Benaim) to conclude.

Synchronous Q-learning:

1. Replace conditional average $\sum_j p(j|i,u) \min_a Q^n(j,a)$ by evaluation at an actual simulated sample:

$$\min_a Q^n(\zeta_{i,u}(n+1), a),$$

   where $\zeta_{i,u}(n+1) \approx p(\cdot|i,u)$.

2. replace 'full move' by an incremental move, i.e., a convex combination of the previous iterate and the correction term due to the new observation.

The algorithm is:

$$\begin{aligned} Q^{n+1}(i,u) &= (1-a(n))Q^n(i,u) \\ &\quad + a(n)[c(i,u) + \beta \min_{u'} Q^n(\xi_{i,u}(n+1),u')] \\ &= Q^n(i,u) + a(n)[c(i,u) \\ &\quad + \beta \min_{u'} Q^n(\xi_{i,u}(n+1),u') - Q^n(i,u)]. \end{aligned}$$

Limiting ODE is

$$\dot{x}(t) = G(x(t)) - x(t)$$

has the desired Q as its globally asymptotically stable equilibrium ($\|x - Q\|_\infty$ works as a Liapunov function)

$\Longrightarrow$ a.s. convergence to Q

(stability is separately proved).

Asynchronous version (single simulation case):

$$Q^{n+1}(i,u) = Q^n(i,u) + a(n) I\{X_n = i, Z_n = u\} \times [c(i,u) + \beta \min_{u'} Q^n(X_{n+1}, u') - Q^n(i,u)].$$

Limiting ODE: $\dot{x}(t) = \Lambda(t)(G(x(t)) - x(t))$,

$\Lambda(\cdot)$ diagonal, non-negative ('relative frequency')

Convergence to Q if diagonal elements of $\Lambda(\cdot)$ are bounded away from zero

$\Longleftrightarrow$ all pairs (i,u) are sampled comparably often.

('infinitely often' suffices (Yu-Bertsekas))

Problem: slow!

Non-incremental Q-learning

Fix $N :=$ number of samples per stage. The algorithm is:

$$Q^{n+1}(i,u) = c(i,u) + \beta \left(\frac{1}{N} \sum_{m=1}^{N} \min_a Q^n(\xi_{i,u}^m(n+1), a) \right),$$

where:

- $\{\xi_{i,u}^m(n)\}$ are IID $\approx p(\cdot|i,u)$ for each (i,u), and,
- $\{\xi_{i,u}^m(n)\}_{i,u,m,n}$ are independent.

This is equivalent to

$$Q^{n+1}(i,u) = c(i,u) + \beta \sum_j \tilde{p}^{(n)}(j|i,u) \min_a Q^n(j,a),$$

where $\tilde{p}^{(n)}(\cdot|i,u)$ are the *empirical transition probabilities* given by

$$\tilde{p}^{(n)}(j|i,u) := \frac{1}{N} \sum_{m=1}^{N} I\{\xi_{i,u}^m(n+1) = j\}.$$

For a fixed sample run, we can view this as 'quenched' randomness, leading to a time-dependent sequence of transition matrices.

Claim: $Q^n \to Q$ a.s.!

Empirical observation: Convergence extremely fast initially to a 'ball park' estimate, then very slow.

$\Longrightarrow$ one can consider hybrid schemes where one switches to stochastic approximation after the initial period.

Idea of proof

Consider a controlled Markov chain $\{X_n\}$ governed by *time-inhomogeneous* transition probabilities

$$\tilde{p}^{(n)}(j|i,u), n \geq 0.$$

V^n in value iteration (always) has the interpretation of being the optimal finite horizon cost with 'terminal cost' V^0, i.e.,

$$V^n(i) = \min_{\{Z_n\}} E\left[\sum_{m=0}^{n-1} \beta^m c(X_m, Z_m) + \beta^n V^0(X_n) | X_0 = i\right]$$

Thus

$$V^n(i) = E\left[\sum_{m=0}^{n-1} \beta^m c(X_m^*, v^*(m, X_m^*)) + \beta^n V^0(X_n^*) | X_0^* = i\right],$$

where $(X_n^*, v^*(n, X_n^*))$ is the optimal state-control process, defined consistently because the function $v(n, \cdot)$ depends on the remaining time horizon.

Similarly,

$$Q^n(i, u) = E\left[\sum_{m=0}^{n-1} \beta^m c(X_m^*, Z_m^*) + \beta^n \min_a Q^0(X_n^*, a) | X_0^* = i\right],$$

where $Z_0^* = u$ and $Z_n^* = v^*(n, X_n^*)$ thereafter.

Consider time-reversed version of this:

$$Q^n(i,u) = E\left[\sum_{m=-n}^{-1} \beta^m c(X_m^*, Z_m^*) + \beta^n \min_a Q^0(X_0^*, a) | X_0^* = i\right].$$

For each $i, u, -n$, generate a chain from i, u.

Consider iterates $Q^m, \check{Q}^m, m \geq -n$, initiated at $(i,u), (i',u')$ resp., and associated state-control processes (X_n^*, Z_n^*), $(\hat{X}_n^*, \hat{Z}_n^*)$.

Fact: As $n \uparrow \infty$, $X_n^*, \hat{X}_n^*$ couple a.s.

(Needs a suitable irreducibility & aperiodicity hypothesis.)

Fact: As $n \uparrow \infty$, $X_n^*, \hat{X}_n^*$ couple a.s.

(Recall Propp-Wilson scheme for exact sampling according to the stationary distribution of a Markov chain through backward coupling.)

$\Longrightarrow Q^n(i,u) - \breve{Q}^n(i',u')$ converges a.s.

But

$$\begin{aligned} Q^n(i,u) \;=\; & c(i,u) + \beta \sum_j \tilde{p}^{(n)}(j|i,u)(\min_a Q^n(j,a) - \\ & \breve{Q}^n(i,u)) + \beta \breve{Q}^n(i,u) \end{aligned}$$

Iterating, one gets

$$
\begin{aligned}
& Q^n(i,u) \\
&= c(i,u) + \beta \sum_j \tilde{p}^{(n)}(j|i,u)(\min_a(Q^n(j,a) - \breve{Q}^n(i,u)) \\
&\quad + \beta(c(i,u) + \beta \sum_j \tilde{p}^{(n)}(j|i,u)(\min_a(Q^n(j,a) - \breve{Q}^n(i,u))))) \\
&\quad \ldots\ldots \\
&= c(i,u) \sum_{m=0}^{n} \beta^m + \\
&\quad \beta \sum_{m=0}^{n} \beta^m (\sum_j \tilde{p}^{(n-m)}(j|i,u)(\min_a(Q^{n-m}(j,a) - \\
&\quad \breve{Q}^{n-m}(i,u))) + \beta^{n+1} Q^0(i,a).
\end{aligned}
$$

By coupling argument, the second term on right converges a.s. Hence $Q^n(i,u) \to Q^*(i,u)$ a.s.

Blackwell-Dubins lemma: $\{Y_n\}$ bounded, $Y_n \to Y$ a.s, $\{\mathcal{F}_n\}$ nested and either $\uparrow$ or $\downarrow \mathcal{F}$. Then a.s.,

$$E[Y_n|\mathcal{F}_n] \to E[Y|\mathcal{F}].$$

Thus $Q^{n+1} - Q^n \to 0$ a.s.

$\Longrightarrow E[Q^{n+1}|\mathcal{F}_n] - Q^n \to 0$ a.s.

$\Longrightarrow$

$E[c(i,u) + \beta \sum_j \tilde{p}^{(n)}(j|i,u) \min_b Q^n(j,b)|\mathcal{F}_n] - Q^n(i,u) \to 0$ a.s.

$\Longrightarrow c(i,u) + \beta \sum_j p(j|i,u) \min_b Q^n(j,b) - Q^n(i,u) \to 0$ a.s.

$\Longrightarrow Q^*$ satisfies the DP equation

$\Longrightarrow Q^* = Q$.

Trade-off: larger $N \Longrightarrow$ faster convergence and less fluctuations, but higher computation per iterate.

Future work:

- asynchronous version
- sample complexity

  (some progress achieved in both)

- other cost criteria

- more general state spaces

- *function approximation*

“With every mistake we must
surely be learning,
still my guitar gently weeps.”

– George Harrison