

Reinforcement Learning

An Introductory Note

Jingye Wang

✉ wangjy5@shanghaitech.edu.cn

Spring 2020

Contents

1	Introduction	3
2	Review of Basic Probability	5
2.1	Interpretation of Probability	5
2.2	Transformations	5
2.3	Limit Theorem	5
2.4	Sampling & Monte Carlo Methods	6
2.5	Basic Inequalities	8
2.6	Concentration Inequalities	10
2.7	Conditional Expectation	12
3	Bandit Algorithms	14
3.1	Bandit Models	14
3.2	Stochastic Bandits	14
3.3	Greedy Algorithms	15
3.4	UCB Algorithms	16
3.5	Thompson Sampling Algorithms	17
3.6	Gradient Bandit Algorithms	18
4	Markov Chains	20
4.1	Markov Model	20
4.2	Basic Computations	20
4.3	Classifications	21

CONTENTS	2
4.4 Stationary Distribution	22
4.5 Reversibility	22
4.6 Markov Chain Monte Carlo	23
5 Markov Decision Process	25
5.1 Markov Reward Process	25
5.2 Markov Decision Process	26
5.3 Dynamic Programming	28
6 Model-Free Prediction	33
6.1 Monte-Carlo Policy Evaluation	33
6.2 Temporal-Difference Learning	35
7 Model-Free Control	37
7.1 On Policy Monte-Carlo Control	37
7.2 On Policy Temporal-Difference Control: Sarsa	39
7.3 Off-Policy Temporal-Difference Control: Q-Learning	40
8 Value Function Approximation	41
8.1 Semi-gradient Method	41
8.2 Deep Q-Learning	43
9 Policy Optimization	46
9.1 Policy Optimization Theorem	46
9.2 REINFORCE: Monte-Carlo Policy Gradient	49
9.3 Actor-Critic Policy Gradient	51
9.4 Extension of Policy Gradient	52

4 Markov Chains

The large part of this section was done with references [4, 2, 3].

4.1 Markov Model

Stochastic Processes: A stochastic process is a collection, or, a sequence of random variables $\{X_t, t \in \mathcal{T}\}$. The set \mathcal{T} is the index set of the process. All the r.v.s are defined on a common state space \mathcal{S} .

Markov Property: Given a stochastic process $X_0, X_1, X_2, \dots, X_n$ taking values in the state space \mathcal{S} , the future evolution of the process is independent of the past evolution of the process, i.e.,

$$P(X_{n+1} = j | X_n = i_n, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = P(X_{n+1} = j | X_n = i_n).$$

The above equation holds for the first order Markov property. For the second order Markov property we have $P(X_{n+1} | X_n, \dots, X_0) = P(X_{n+1} | X_n, X_{n-1})$, etc. With Markov property, many calculation tasks can be simplified greatly.

Markov Chain/Process: A sequence of random variables X_0, X_1, X_2, \dots taking values in the state space \mathcal{S} is called a Markov chain if it has Markov property. A Markov process is the continuous-time version of a Markov Chain.

Transition Matrix: For a Markov Chain, let $q_{ij} = P(X_{n+1} = j | X_n = i)$ be the transition probability from state i to state j . Then the matrix Q is called the transition matrix of the chain.

Transition Matrix is a common way to express a Markov chain. Besides that, Markov chain can be represented in a graphical form.

4.2 Basic Computations

With a little abuse of notation, I would use Q to denote the transition matrix when we talk about Markov chain, and $q_{i,j}^n$ to denote the entry $(Q^n)_{i,j}$. Now we introduce some useful computations.

n -step Transition Probability: For a Markov chain, the n -step transition probability from i to j is the probability of being at j exactly n steps after being at i , and

$$P(X_{n+m} = j | X_m = i) = q_{i,j}^n.$$

Proof:

For a Markov chain, the states are time-homogeneous. Thus we have

$$P(X_{n+m} = j | X_m = i) = P(X_n = j | X_0 = i).$$

Hence it follows that

$$\begin{aligned}
 P(X_n = j | X_0 = i) &= \sum_k P(X_n = j, X_{n-1} = k | X_0 = i) && \text{(by LOTP)} \\
 &= \sum_k P(X_n = j | X_{n-1} = k, X_0 = i) P(X_{n-1} = k | X_0 = i) \\
 &= \sum_k q_{k,j} P(X_{n-1} = k | X_0 = i), && \text{(by Markov Property)}
 \end{aligned}$$

then by *induction* from 2 to $n - 1$, we have

$$P(X_{n+m} = j | X_m = i) = q_{i,j}^n.$$

□

With n -step transition probability, we have the *Chapman-Kolmogorov Equation*.

Chapman-Kolmogorov Equation: For $m, n \geq 0$, we have

$$P(X_{m+n} = j | X_0 = i) = \sum_k P(X_m = k | X_0 = i) P(X_n = j | X_0 = k).$$

The equation can be proved by matrix identity that $q_{i,j}^{m+n} = \sum_k q_{i,k}^m q_{k,j}^n$. By the equation, for a Markov chain with transition matrix P , the Markov property can be generalized to

$$P(X_{n+1} = j | X_{n-m} = i, X_{n-m-1} = i_{n-m-1}, \dots, X_0 = i_0) = P(X_{n+1} = j | X_{n-m} = i) = q_{i,j}^{m+1}$$

for $m < n$ and $m \geq 0$.

4.3 Classifications

Depending on whether they are visited over and over again in the long run or are eventually abandoned, the states of a Markov chain can be classified as recurrent or transient.

Recurrent and Transient states: State i of a Markov chain is recurrent if starting from i , the chain can always return to i . Otherwise, the state is transient, which means that if the chain starts from i , there is a positive probability of never returning to i .

Irreducible and Reducible Chain: A Markov chain with transition matrix Q is irreducible if for any two states i and j , it is possible to go from i to j in a finite number of steps (with positive probability). That is, for any states i, j there is some positive integer n such that the (i, j) entry of Q^n is positive. A Markov chain that is not irreducible is called reducible.

In an irreducible Markov chain with a finite state space, all states are recurrent.

Period: For a Markov chain with transition matrix Q , the period of state i , denoted $d(i)$, is the greatest common divisor of the set of possible return times to i . That is,

$$d(i) = \gcd\{n > 0 \mid q_{i,i}^n > 0\}.$$

If $d(i) = 1$, state i is said to be aperiodic. If the set of return times is empty, set $d(i) = +\infty$.

4.4 Stationary Distribution

Stationary Distribution: A row vector $\mathbf{s} = (s_1, \dots, s_M)$ such that $\sum_i s_i = 1$ is a stationary distribution for a Markov chain with transition matrix Q if

$$\sum_i s_i q_{i,j} = s_j$$

for all j .

Any irreducible Markov chain has a unique stationary distribution.

Doubly Stochastic Matrix: A nonnegative matrix such that the row sums and the column sums are all equal to 1 is called a doubly stochastic matrix.

If the transition matrix Q of a Markov chain is a doubly stochastic matrix, then the uniform distribution over all states, $(1/M, 1/M, \dots, 1/M)$, $M = |\mathcal{S}|$, is a stationary distribution of the chain.

Convergence to Stationary Distribution: Let X_0, X_1, \dots be a Markov chain with stationary distribution \mathbf{s} and transition matrix Q , such that some power Q^m is positive in all entries. (These assumptions are equivalent to assuming that the chain is irreducible and aperiodic.) Then $P(X_n = i)$ converges to s_i as $n \rightarrow \infty$. In terms of the transition matrix, Q^n converges to a matrix in which each row is \mathbf{s} .

Ergodic Markov chain: A Markov chain is called ergodic if it is irreducible, aperiodic, and all states have finite expected return times (positive recurrent).

For an ergodic Markov chain X_0, X_1, \dots , there exists a unique stationary distribution π , which is the limiting distribution of the chain. That is

$$\pi_j = \lim_{n \rightarrow \infty} q_{i,j}^n, \forall i, j.$$

4.5 Reversibility

Reversibility: Let Q be the transition matrix of a Markov chain. Suppose there is $\mathbf{s} = (s_1, \dots, s_M)$ with $s_i \geq 0$, $\sum_i s_i = 1$, such that

$$s_i q_{i,j} = s_j q_{j,i}$$

for all pairs of states i and j .

This equation is called the reversibility or detailed balance condition. We say that the chain is reversible with respect to \mathbf{s} if it holds, and such \mathbf{s} is a stationary distribution of the chain.

Detailed Balance Equation: *If for an irreducible Markov chain with transition matrix Q , there exists a probability solution π to the detailed balance equations*

$$\pi_i q_{i,j} = \pi_j q_{j,i}$$

for all pairs of states i and j , then this Markov chain is positive recurrent, time-reversible and the solution π is the unique stationary distribution.

4.6 Markov Chain Monte Carlo

Monte Carlo method is a simulation approach where we generate random values to approximate a quantity. A basic form of such method is directly generating *i.i.d.* draws X_1, X_2, \dots, X_n from a given distribution, then by the law of large numbers we can make a desired approximate if n is large. However, staring at a density function does not immediately suggest how to get a random variable with that density.

Fortunately, for this limitation, we have *Markov chain Monte Carlo* (MCMC), a powerful collection of algorithms, to enable us to simulate from complicated distributions using Markov chains. The basic idea is to *build your own Markov chain* so that the distribution of interest is the stationary distribution of the chain.

Convergence to stationary distribution: *Let X_0, X_1, \dots be a Markov chain with stationary distribution s and transition matrix Q , such that the chain is irreducible and aperiodic. Then $P(X_n = i)$ converges to s_i as $n \rightarrow \infty$.*

With the above theorem, which actually has been mentioned in *ergodic Markov chain*, we can approach the desired s by running our chain for a long time.

Metropolis-Hastings: Metropolis-Hastings allows us to start with any *irreducible* Markov chain on the state space of interest and then modify it into a new Markov chain that has the desired stationary distribution.

Recall: In an irreducible Markov chain, for any two states i and j it is possible to go from i to j in a finite number of steps.

Metropolis-Hastings Algorithm: *Let $s = (s_1, \dots, s_M)$ be a desired stationary distribution on state space. Suppose that $Q = q_{ij}$ is the transition matrix for any irreducible Markov chain on state space $\{1, \dots, M\}$. Then we can use a chain with transition matrix Q to construct a collection of states sample X_0, X_1, \dots with stationary distribution s .*

Algorithm 5 Metropolis-Hastings

```

1: input the desired distribution  $s = (s_1, \dots, s_M)$ ; the chain with transition matrix  $Q$ ; the initial state  $X_0$ ;
2: for  $n = 0, 1, \dots$  do: # assume that  $X_n = i$ ;
3:   Sample the next state  $j$  according to  $Q$ ;
4:   Calculate the acceptance probability  $a_{ij} = \min\left(\frac{s_j q_{ji}}{s_i q_{ij}}, 1\right)$ ;
5:    $X_{n+1} \leftarrow \begin{cases} j & \text{with probability } a_{ij}; \\ i & \text{with probability } 1 - a_{ij}; \end{cases}$ 
6: end for

```

In practice, a useful trick is *Burn-in*, which discards the initial iterations and retains X_m, X_{m+1}, \dots for some m . The key of the algorithm is that the moves are proposed according to the original chain, but the proposal may or may not be accepted. By the *reversibility condition*, it can be showed that the sequence X_0, X_1, \dots constructed by the Metropolis-Hastings algorithm is a *reversible* Markov chain with stationary distribution s .

Gibbs Sampler: Gibbs sampling is an MCMC algorithm for obtaining approximate draws from a joint distribution, based on sampling from conditional distributions one at a time: at each stage, one variable is updated (keeping all the other variables fixed) by drawing from the conditional distribution of that variable given all the other variables.

Gibbs Sampling Algorithm: Let X and Y be discrete r.v.s with joint PMF $p_{X,Y}(x, y) = P(X = x, Y = y)$. We wish to construct a two-dimensional Markov chain (X_n, Y_n) whose stationary distribution is $p_{X,Y}$. The systematic scan Gibbs sampler proceeds by updating the X -component and the Y -component in alternation. If the current state is $(X_n, Y_n) = (x_n, y_n)$, then we update the X -component while holding the Y -component fixed, and then update the Y -component while holding the X -component fixed.

Algorithm 6 Gibbs Sampling

```

1: input the desired joint distribution  $P(X, Y)$ ; the initial state  $X_0, Y_0$ ;
2: for  $n = 0, 1, \dots$  do:
3:   Sample the next state  $X_{n+1}$  from  $P(X, Y = Y_n)$ ;
4:   Sample the next state  $Y_{n+1}$  from  $P(X = X_{n+1}, Y)$ ;
5: end for

```

The algorithm can be generalized to high dimensional easily in the light of line 3 and 4.