可数状态空间的平均成本马氏决策过程

张俊玉[1], 吴怡婷[1], 夏 俐[2], 曹希仁[3†]

(1. 中山大学 数学学院, 广东 广州 510275; 2. 中山大学 管理学院, 广东 广州 510275; 3. 香港科技大学 电子与计算机工程系, 中国 香港)

摘要: 具有可数状态空间的马尔可夫决策过程(Markov decision process, MDP)在平均准则下, 最优(平稳)策略不一定存在. 本文研究平均准则可数状态MDP中满足最优不等式的最优策略. 不同于消去折扣(因子)方法, 利用离散的Dynkin公式推导本文的主要结果. 首先给出遍历马氏链的泊松方程和两个零常返马氏链的例子, 证明了满足两个方向相反的最优不等式的最优策略存在性. 其次, 通过两个比较引理和性能差分公式, 证明了正常返链和多链最优策略的存在性, 并进一步推广到其他情形. 特别地, 本文通过几个应用举例, 说明平均准则性能敏感的本质. 本文的结果完善了可数状态MDP在平均准则下的最优不等式的理论.

关键词: 马尔可夫决策过程; 平均准则; 可数状态空间; Dynkin公式; 泊松方程; 性能敏感

引用格式: 张俊玉, 吴怡婷, 夏俐, 等. 可数状态空间的平均成本马氏决策过程. 控制理论与应用, 2021, 38(11): 1707 – 1716

DOI: 10.7641/CTA.2021.10763

Average cost Markov decision processes with countable state spaces

ZHANG Jun-yu[1], WU Yi-ting[1], XIA Li[2], CAO Xi-ren[3†]

(1. School of Mathematics, Sun Yat-Sen University, Guangzhou Guangdong 510275, China;
2. School of Business, Sun Yat-Sen University, Guangzhou Guangdong 510275, China;
3. Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong, China)

Abstract: For the long-run average of a Markov decision process (MDP) with countable state spaces, the optimal (stationary) policy may not exist. In this paper, we study the optimal policies satisfying optimality inequality in a countable-state MDP under the long-run average criterion. Different from the vanishing discount approach, we use the discrete Dynkin's formula to derive the main results of this paper. We first provide the Poisson equation of an ergodic Markov chain and two instructive examples about null recurrent Markov chains, and demonstrate the existence of optimal policies for two optimality inequalities with opposite directions. Then, from two comparison lemmas and the performance difference formula, we prove the existence of optimal policies under positive recurrent chains and multi-chains, which is further extended to other situations. Especially, several examples of applications are provided to illustrate the essential of performance sensitivity of the long-run average. Our results make a supplement to the literature work on the optimality inequality of average MDPs with countable states.

Key words: Markov decision process; long-run average; countable state spaces; Dynkin's formula; Poisson equation; performance sensitivity

Citation: ZHANG Junyu, WU Yiting, XIA Li, et al. Average cost Markov decision processes with countable state spaces. *Control Theory & Applications*, 2021, 38(11): 1707 – 1716

1 Introduction

The Markov decision process (MDP) is an important optimization theory for stochastic dynamic systems and has wide applications; see, e.g. [1–10]. In this paper, we study the long-run average MDP with a countable state space. It is well known that in this problem an optimal policy may not exist, and if it exists, it may be history-dependent, not necessarily a stationary (or else called a Markov) policy.

In the literature, conditions have been found that guarantee the existence of the long-run average optimal policy for MDPs with countable states [7–8,10–12]. The existence conditions are usually stated in terms of discounted value functions with the discount factor approaching one. It is also proved that there are average optimal policies for MDPs with countable states for

Received 20 August 2021; accepted 25 October 2021.
†Corresponding author. E-mail: eecao@ust.hk.
Recommended by Associate Editor: ZHAO Qian-chuan.
Supported by the National Natural Science Foundation of China (61673019, 61773411, 11931018, 62073346), the Guangdong Province Key Laboratory of Computational Science at the Sun Yat-sen University (2020B1212060032) and the Guangdong Basic and Applied Basic Research Foundation (2021A1515010057, 2021A1515011984).

which an optimality inequality, instead of an equality, holds. The approach is presented in details in [10], and it is called the "vanishing discount approach" in [7], and the "differential discounted reward" approach in [8].

In this paper, we use the discrete Dynkin's formula (e.g. page 122 in [13]) to derive the average optimal policy. We observed that the long-run average of a countable MDP does not depend on its value at any finite state, or at any "non-frequently visited" states, which is an important fact studied in the literature [3–4, 6, 8, 14]. We construct several examples about null recurrent Markov chains, which motivate the derivation of the strict optimality inequality and optimality equality. We present two comparison lemmas and the performance difference formula via the discrete Dynkin's formula, which is an important tool to prove the existence of average optimal policies.

The Poisson equation provides a method to solve the long-run average cost, where the existence and properties of its solutions are studied in many literature, for example, refer to page 269 to 303 in [5] which considers the Poisson equation with time-homogeneous Markov chains on a countable state space. Actually, the Poisson equation is a different view to study the existence of the long-run average optimal policy. From the Poisson equation, we derive the optimality inequality/equality and the existence of average optimal policies of positive recurrent chains, multi-chains and other generalized forms such as considering the performance sensitivity. One of the optimality inequality directions is presented as the sufficient condition for average optimal. The more general optimality equations are obtained by the approach of performance difference formulas, which is a useful approach beyond the dynamic programming and has been successfully applied to many problems (e.g. [15–16]).

The remainder of this paper is organized as follows. In Section 2, we introduce the optimization problem and derive the Poisson equation of ergodic Markov chains. In Section 3, we give two examples to demonstrate the existence of the long-run average optimal policy for two optimality inequalities with opposite directions. Our main results are given in Section 4, where the existence of optimal policies under multi-chains, null and positive recurrent chains are provided and some examples and applications are also presented. Finally, we conclude this paper in Section 5.

2 Preliminaries

2.1 The problem

Consider a stochastic chain $\boldsymbol{X} := \{X_0, X_1, \cdots\}$ with a countable state space $\mathcal{S} = \{0, 1, 2, \cdots\}$, where X_k is the system state at time k. Let $S := |\mathcal{S}|$ be the number of states in $\mathcal{S}$; S may be infinity. Let $P := [P_{i,j}]_{i,j=1}^{S}$ be the transition probability matrix, which may depend on the history of $\boldsymbol{X}$, denoted as $\hbar_k := \{X_0, X_1, \cdots, X_k\}$. If $P_{i,j}$ depends only on the current state $X_k = i$, then $\boldsymbol{X}$ is a Markov chain. The states of a Markov chain may be classified into transient states, and null or positive recurrent states. Let $(\Omega, \mathcal{F}, \mathcal{P})$ be the probability space generated by the chain, and E be the corresponding expectation.

In optimization problems, at any state $i \in \mathcal{S}$, we take an action $\alpha \in \mathcal{A}(i)$, with $\mathcal{A}(i)$ being the set of all available actions at i. The action α determines the transition probabilities at i. The action α can be determined by a policy d. We use a superscript to denote the policy or action associated with a quantity, e.g. $P_{i,j}^{d}$ or $P_{i,j}^{\alpha}$, $i, j \in \mathcal{S}$. The stochastic chain is denoted by $\boldsymbol{X}^d := \{X_0^d, X_1^d, \cdots\}$. A policy may depend on the history $\hbar_k = \{x_0, a_0, x_1, a_1, \cdots, x_k\}$ ($x_i \in \mathcal{S}, i = 0, 1, \cdots, k; a_i \in \mathcal{A}(x_i), i = 0, 1, \cdots, k-1$), denoted by $d_{\hbar_k}$, and, if necessary, the corresponding quantities are denoted by a subscript $\hbar_k$, e.g. $P_{\hbar_k}^{d}$ [1]. If the policy d depends only on the current state, then d is called a Markov policy. Let $\mathit{\Pi}$ be the space of all policies, including the history-dependent policies, and $\mathit{\Pi}_0 \subset \mathit{\Pi}$ be the space of all Markov policies. For each history-dependent randomized policy, we can construct a randomized Markov policy with the same joint probablity distribution of states and actions. And for most Markov decision problems, we can consider deterministic Markov policy instead of randomized Markov policy, see, e.g. [8]. So we mainly consider deterministic policy in the followings.

At state i with action α determined by policy d, a cost (or reward), denoted by $C^{\alpha}(i)$ or $C^{d}(i)$. The discounted cost criterion with discount factor $0 < \beta < 1$ under policy d is

$$\eta_{\beta}^{d}(i) = \mathrm{E}^{d}\{\sum_{k=0}^{\infty} \beta^k C^{d}(X_k^d) | X_0^d = i\}.$$

The long-run average cost criterion under policy d is

$$\eta^{d}(i) = \limsup_{N\to\infty} \frac{1}{N} \mathrm{E}^{d}\{\sum_{k=0}^{N-1} C^{d}(X_k^d) | X_0^d = i\}. \quad (1)$$

The optimal performance of long-run average cost is $\eta^*(i) = \inf\limits_{d\in\mathit{\Pi}} \eta^{d}(i)$. The goal of optimization is to find an optimal policy d^*, if it exists, that attains the optimal performance [2]

$$d^* = \arg\{\min_{d\in\mathit{\Pi}} \eta^{d}(i),\ i \in \mathcal{S}\}.$$

2.2 The Poisson equation

Consider a Markov chain $\{X_k, k = 0, 1, \cdots\}$ under a Markov policy d (We omit the superscript "d", for a generic discussion). Let $\tau(i, j) = \min\{t > 0 : X_t =$

[1] Instead of $P^{d_{\hbar_k}}$, with a double superscript.

[2] Under some technical conditions, such a policy indeed exists.

$j|X_0 = i\}$ be the first passage time from i to j. For an ergodic Markov chain, the stationary probability distribution exists, the long-run average η does not depend on the initial state i, and $\tau(i,j) < \infty$ for all $i, j \in \mathcal{S}$. Furthermore, we have $\eta = \lim\limits_{k\to\infty} \mathrm{E}[C(X_k)|X_0 = i]$. We further define

$$g(i,j) := \mathrm{E}\{\sum_{k=0}^{\tau(i,j)-1} [C(X_k) - \eta]|X_0 = i\},\ i, j \in \mathcal{S}.$$

First, we choose a reference state, e.g. state 0, and define

$$g(i) := g(i,0) = \mathrm{E}\{\sum_{k=0}^{\tau(i,0)-1} [C(X_k) - \eta]|X_0 = i\},\ i \in \mathcal{S}, \quad (2)$$

with $g(0) = g(0,0) = 0$. $g(i), i = 0, 1, \cdots$, is called a *potential function* of the Markov chain[3]. And we further assume $|g(i)| < \infty,\ i \in \mathcal{S}$.

Let $A := P - I$ be the discrete version of the infinitesimal generator, with I being the identity matrix. $g := (g(0)\ g(1)\ \cdots)^{\mathrm{T}}$, $C := (C(0)\ C(1)\ \cdots)^{\mathrm{T}}$, and $e := (1\ 1\ \cdots)^{\mathrm{T}}$.

Lemma 1 The potential function $g(i)$ of a Markov chain, $i \in \mathcal{S}$, in (2) satisfies the Poisson equation $Ag + C = \eta e$.

Proof From (2), we have

$$\begin{aligned} g(i) &= C(i) - \eta + \mathrm{E}\{\mathrm{E}[\sum_{k=1}^{\tau(i,0)-1} [C(X_k) - \eta]|X_1]|X_0 = i\} = \\ &C(i) - \eta + \mathrm{E}\{\mathrm{E}[\sum_{k=1}^{\tau(X_1,0)} [C(X_k) - \eta]|X_1]|X_0 = i\} = \\ &C(i) - \eta + \mathrm{E}\{g(X_1)|X_0 = i\},\ i \in \mathcal{S}. \end{aligned}$$

On every sample path with $X_1 \neq 0$ (when $X_1 = 0$, Lemma 1 is easy to verify since $\tau(i,0) = 1, i \in \mathcal{S}$), the Markov chain reaches state 0 from state X_1 at time 1 after time $\tau(i,0)-1$, i.e., $\tau(X_1,0) = \tau(i,0)-1$, and the second equality in the above equation holds. Thus, this equation is the Poisson equation

$$Ag(i) + C(i) = \eta,\ i \in \mathcal{S}, \quad (3)$$

in which Ag is a vector, and $Ag(i)$ is its ith component. □

For multi-chains, η depends on i and the Poisson equation is[4]:

$$Ag(i) + C(i) = \eta(i),\ i \in \mathcal{S}.$$

Obviously, the potential function is only up to an additive constant; i.e., if $g(i), i \in \mathcal{S}$, is a solution to the Poisson equation, so is $g(i) + c$ for any constant c. Any state $i \in \mathcal{S}$ can be chosen as a reference state and we may set $g(i) = 0$.

Remark 1 For null recurrent Markov chains, $\tau(i,j)$ may be infinity, and $g(i)$ in (2) is not well defined. However, in some special cases, there might be some function $g(i)$ and a constant η such that the Poisson equation (3) holds, see Example 1.

3 Examples

To motivate our further research, let us first consider some examples.

Example 1 This is a well-known example [6–8, 10] that shows there is an optimal policy for which the optimality equation for finite chains does not hold, instead, it satisfies an inequality. As in the literature, we use the discounted performance to approach the long-run average.

The state space is $\mathcal{S} = \{0, 1, 2, \cdots\}$. At state $i \geqslant 1$, there is a "null" action with $P_{i,i-1} = 1$ and cost $C(i) = 1$. At state 0, there are actions a and b, with $C^a(0) = 0$ and $C^b(0) = 1$. The transition probabilities at state 0 for both actions are the same as $P_{0,i} = p_i$, $i \geqslant 1$, where p_i, $i \geqslant 1$, be a probability distribution on $i \geqslant 1$, with $\sum\limits_{i=1}^{\infty} ip_i = \infty$.

Let f (respectively, d) be the stationary policy that chooses a (respectively, b) when at state 0. The costs under d are identically 1, and hence $\eta^d(i) = 1$ for all i. The Markov chain under policy f is null recurrent, and the number of cost 1 is higher than that of cost 0, so we have $\eta^f(i) = 1$ for all i. In both cases, the limit in the long-run average (1) exists, and both d and f are optimal long-run average policies (among two policies $\Pi := \{f, d\}$).

Next, we have $\eta_\beta^d(i) = \dfrac{1}{1-\beta}$, and because $C^a(0) = 0$, it is clear that $\eta_\beta^f(i) < \dfrac{1}{1-\beta}$. Thus, f is discount optimal for all $\beta \in (0, 1)$.

Now, we focus on policy f and suppress the superscript "f". We have $\eta(i) \equiv \eta = 1$ for all i, and by the structure of the Markov chain under f, we have $\eta_\beta(i) = \dfrac{1-\beta^i}{1-\beta} + \beta^i \eta_\beta(0), i \geqslant 0$. Therefore, for $i \geqslant 0$,

$$\begin{aligned} g_\beta(i) &:= \eta_\beta(i) - \eta_\beta(0) = \\ &(\frac{1-\beta^i}{1-\beta})[1 - (1-\beta)\eta_\beta(0)] = \\ &(1 + \beta + \cdots + \beta^{i-1})[1 - (1-\beta)\eta_\beta(0)]. \end{aligned} \quad (4)$$

[3] In the literature, the solution to a Poisson equation is called a potential function; the conservative law for potential energy holds, see [17] and [4].

[4] The theory for multi-class decomposition and optimization for finite Markov chains is well developed, see [4, 8] for homogeneous chains, and [3] for nonhomogeneous chains. The decomposition of countable state chains is similar to that for the finite chains, except that there are infinitely many sub-chains.

The Poisson equation for the discounted problem is

$$\eta_\beta(i) = C^\alpha(i) + \beta \sum_j P^\alpha_{i,j}\eta_\beta(j),\ i \in \mathcal{S},$$

where α is the action taken at state i. We have

$$\eta_\beta(i) - \eta_\beta(0) = \\ C^\alpha(i) - (1-\beta)\eta_\beta(0) + \beta \sum_j P^\alpha_{i,j}[\eta_\beta(j) - \eta_\beta(0)].$$

With $g_\beta(i) = \eta_\beta(i) - \eta_\beta(0)$, we have

$$g_\beta(i) + (1-\beta)\eta_\beta(0) = C^\alpha(i) + \beta \sum_j P^\alpha_{i,j} g_\beta(j). \tag{5}$$

It has been shown (e.g. [10]) that $\lim\limits_{\beta\to 1}(1-\beta)\eta_\beta(i) = \eta = 1$ for all $i \in \mathcal{S}$. From (4), we have

$$g(i) := \lim_{\beta\to 1} g_\beta(i) = i(1-\eta) = 0,\ i \in \mathcal{S}.$$

However, this convergence is not uniform on $\mathcal{S}$ as $\beta \to 1$.

When $i > 0$, there is only one term, $P_{i,i-1} = 1$, in the summation of the right-hand-side of (5), so we may take the limit of $\beta \to 1$ on both sides and obtain

$$g(i) + \eta = C(i) + g(i-1),\ i > 0.$$

Indeed, we have $0 + 1 = 1 + 0$. However, when $i = 0$, the summation on the right-hand-side of (5) contains infinitely many terms and we cannot exchange the order of $\lim\limits_{\beta\to 1}$ and $\sum\limits_j$ (Fatou's lemma for $\liminf$, non-uniform for $\lim$). Indeed, at $i = 0$ we have $0 + 1 > 0 + \sum\limits_j P_{i,j} \times 0$, i.e.

$$g(0) + \eta > C^a(0) + \sum_j P^a_{0,j} g(j) = \\ \min_{\alpha=a,b}\{C^\alpha(0) + \sum_j P^\alpha_{0,j} g(j)\}. \tag{6}$$

This is called the optimality inequality for the long-run average MDP with countable states in the literature.

Example 2 In this example, we show that the direction of the "optimality inequality" (6) can be reversed. Consider the same Markov chain as Example 1, but with a different cost at state $i = 0$. Specifically, there are two actions, a and b, having the same transition probabilities at all states $i \in \mathcal{S}$, and the same costs $C(i) = 1$, for all $i \neq 0$. However, we set $C^a(0) = 2$ and $C^b(0) = 3$. Policy f takes action a and policy d takes action b at $i = 0$.

First, we have $\eta^f_\beta(i) < \eta^d_\beta(i)$. Thus, d is not discount optimal for all $\beta \in (0,1)$. Next, because both Markov chains under f and d are null recurrent, the cost change at one state 0 does not change the long-run average. In general, the long-run average of a stochastic chain does not depend on its values in any finite period or any "zero frequently" visited period [3]. This is called the "under-selectivity" [3, 14]. Thus, we have $\eta^f(i) = \eta^d(i) \equiv \eta = 1$ for all i. Therefore f and d are both long-run average optimal.

Now, we analyze policy f. Similar to Example 1, we have (dropping the superscript "f") $\eta_\beta(i) = \dfrac{1-\beta^i}{1-\beta}$ $+ \beta^i\eta_\beta(0), i \geqslant 0$. Therefore (cf. (4)),

$$g_\beta(i) := \eta_\beta(i) - \eta_\beta(0) = \\ (1 + \beta + \cdots + \beta^{i-1})[1 - (1-\beta)\eta_\beta(0)]. \tag{7}$$

Similarly, $\eta = \lim\limits_{\beta\to 1}(1-\beta)\eta_\beta(i)$ for all i. From (7), we have $g(i) := \lim\limits_{\beta\to 1} g_\beta(i) = i(1-\eta) = 0, i \in \mathcal{S}$. This convergence is also not uniform on $\mathcal{S}$ as $\beta \to 1$.

As in Example 1, when $i > 0$, we have $0 + 1 = 1 + 0$, that is, the optimality equation holds:

$$g(i) + \eta = C(i) + \sum_j P_{0,j} g(j),\ i = 1, 2, \cdots.$$

However, when $i = 0$, we have $0 + 1 < 2 + \sum\limits_j P_{i,j} \times 0$; i.e.,

$$g(0) + \eta < C^a(0) + \sum_j P^a_{0,j} g(j) = \\ \min_{\alpha=a,b}\{C^\alpha(0) + \sum_j P^\alpha_{0,j} g(j)\}. \tag{8}$$

Compared with the optimality inequality (6), the inequality sign is reversed in (8).

Remark 2 1) The two inequalities (6) and (8) are not necessary conditions. Also, an average optimal policy may not be the limiting point of discount optimal policies. In the following, we will see that (8) is a sufficient condition for average optimal, and (6) is neither necessary, nor sufficient.

2) In the examples, the Markov chain is null recurrent. So the probability of visiting each state is zero. Therefore, we may arbitrarily change the cost at any state without changing the long-run average; but it changes the relation in the optimality condition. The inequality (6) is due to the null recurrency of the states and the under-selectivity of the long-run average, it is not an essential property in optimization.

4 Optimization in countable state spaces

4.1 Fundamental results

We assume that the transition probability matrix $P = [P_{i,j}]_{i,j\in\mathcal{S}}$ does not depend on time; i.e., the chain $\boldsymbol{X}$ under a Markov policy is a time-homogeneous Markov chain.

Let $r : \mathcal{S} \to \mathcal{R}$ be a function on $\mathcal{S}$, with $\mathrm{E}[r(X_n)|$ $X_0 = i] < \infty, i \in \mathcal{S}$, $n \geqslant 1$; and we also denote it as a column vector $r := (r(0)\ r(1)\ \cdots\ r(i)\ \cdots)^{\mathrm{T}}$. For time-homogeneous Markov chains, it holds that $\mathrm{E}[r(X_{n+1})|X_n = i] = \mathrm{E}[r(X_1)|X_0 = i]$.

In general, we define an infinitesimal operator $A_{\hbar_n}$: it acts on the function r resulting a function, or a vector, $A_{\hbar_n} r$ (the subscript $\hbar_n$ refers to history dependent), with

$$(A_{\hbar_n} r)(i) = \mathrm{E}[r(X_{n+1})|X_n = i] - r(i),$$

$$i \in \mathcal{S},\ n \geqslant 1. \tag{9}$$

We have $A_{\hbar_n} r(i) = \sum_{j\in\mathcal{S}} P_{i,j}^{\hbar_n} r(j) - r(i), i \in \mathcal{S}$, or $A_{\hbar_n} = P^{\hbar_n} - I$, with I being the identity matrix. From (9), we have

$$\begin{aligned}&\mathrm{E}[A_{\hbar_n} r(X_n)|X_0 = i] = \\ &\mathrm{E}\{\mathrm{E}[r(X_{n+1})|X_n] - r(X_n)|X_0 = i\} = \\ &\mathrm{E}[r(X_{n+1})|X_0 = i] - \mathrm{E}[r(X_n)|X_0 = i].\end{aligned}$$

Adding this equation from $n = 0$ to N, we get the following discrete Dynkin's formula:

$$\begin{aligned}&\sum_{k=0}^{N} \mathrm{E}[A_{\hbar_k} r(X_k)|X_0 = i] = \\ &\mathrm{E}[r(X_{N+1})|X_0 = i] - r(i).\end{aligned} \tag{10}$$

Lemma 2 (Comparison Lemma I) Consider any Markov policy $d \in \Pi_0$ associated with P^d, which generates a Markov chain $\{X_k^d, k = 0, 1, \cdots\}$ with long-run average $\eta^d(i), i \in \mathcal{S}$. Suppose there are a constant J and a function $r(i)$: $\mathcal{S} \to \mathcal{R}$, such that

$$\mathrm{E}^d[r(X_n^d)|X_0^d = i] < \infty,\ i \in \mathcal{S},\ n \geqslant 1, \tag{11}$$

$$\lim_{N\to\infty} \frac{1}{N} \mathrm{E}^d[r(X_N^d)|X_0^d = i] = 0,\ i \in \mathcal{S} \tag{12}$$

and the optimality inequality

$$J + r(i) \leqslant C^\alpha(i) + \sum_{j\in\mathcal{S}} P_{i,j}^\alpha r(j) \tag{13}$$

holds for any $i \in \mathcal{S}$, with $\alpha = d(i)$ being the action taken at state i, then

$$J \leqslant \eta^d(i),\ i \in \mathcal{S}. \tag{14}$$

Proof By (10) and (12), we have (omitting the superscript "d")

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} \mathrm{E}[Ar(X_k)|X_0 = i] = 0.$$

Thus,

$$\begin{aligned}&\eta(i) - J = \\ &\limsup_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} \mathrm{E}[Ar(X_k) + C(X_k) - J|X_0 = i].\end{aligned}$$

The theorem follows directly from this difference formula. □

Remark 3 1) Condition (12) can be replaced by a more general one

$$\lim_{N\to\infty} \frac{1}{N} \mathrm{E}^d[r(X_N^d)|X_0^d = i] \leqslant 0.$$

However, for positive costs, the relation $<$ does not make practical sense.

2) If (13) changes to $J + r(i) \geqslant C^\alpha(i) + \sum_{j\in\mathcal{S}} P_{i,j}^\alpha r(j)$, then (14) becomes $J \geqslant \eta^d(i)$, and (12) can be relaxed to $\lim_{N\to\infty} \frac{1}{N} \mathrm{E}^d[r(X_N^d)|X_0^d = i] \geqslant 0$.

3) The Markov chain may be a multi-chain. When the Markov chain is recurrent, $\eta^d(i) \equiv \eta, i \in \mathcal{S}$, being a constant.

4) The condition (13) is only sufficient, it is not necessary as shown in Example 1. It may not need to hold at some null recurrent states.

Example 3 It is interesting to note that in Example 2, condition (8) satisfies the optimality inequality (13) for both policies d and f, and hence it is a sufficient condition for J to be the optimal average. However, in Example 1, the inequality (6) is not a sufficient condition.

For history-dependent policies, it is more convenient to state the comparison lemma for all policies.

Lemma 3 (Comparison Lemma II) Let $d \in \Pi$ associated with a stochastic chain $\{X_k^d, k = 0, 1, \cdots\}$ and long-run average $\eta^d(i), i \in \mathcal{S}$. The action α determines the transition probabilities $P_{i,j}^\alpha$, $i, j \in \mathcal{S}$. Suppose there is a function $r(i)$: $\mathcal{S} \to \mathcal{R}$, satisfying (11) and (12) for all $d \in \Pi$. If there is a constant J such that the optimality inequality

$$J + r(i) \leqslant \min_{\alpha\in\mathcal{A}(i)} \{C^\alpha(i) + \sum_{j\in\mathcal{S}} P_{i,j}^\alpha r(j)\} \tag{15}$$

holds for any $i \in \mathcal{S}$, then $J \leqslant \eta^d(i), i \in \mathcal{S}$ for all $d \in \Pi$.

Proof For a optimization problem with history dependent randomized policy, we can transform into considering randomized Markov policy, see [8]. So the proof can be derived by Lemma 2. □

The performance difference formula may contain more information than the above two comparison lemmas. For history-dependent chains, the potentials and Poisson equations are not well studied, and the "lim sup" average does not make sense at transient states, so we have to assume that the limit in (1) exists.

Lemma 4 Let $f \in \Pi$ and $d \in \Pi_0$, and g^d is the potential of d. Assume that

$$\lim_{N\to\infty} \frac{1}{N} \mathrm{E}^f[g^d(X_N^f)|X_0^f = i] = 0. \tag{16}$$

Then, we have

$$\begin{aligned}&\eta^f(i) - \eta^d(i) = \\ &\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} \mathrm{E}^f\{\{[P_{\hbar_k}^f g^d](X_k^f) + C^f(X_k^f)\} - \\ &\{[P^d g^d](X_k^f) + C^d(X_k^f)\}|X_0^f = i\} + \\ &\{\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} \mathrm{E}^f[\eta^d(X_k^f)|X_0^f = i] - \eta^d(i)\}.\end{aligned} \tag{17}$$

When $\lim_{N\to\infty} \mathrm{E}^f[\eta^d(X_N^f)|X_0^f = i]$ exists,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} \mathrm{E}^f[\eta^d(X_k^f)|X_0^f = i]$$

in (17) is equal to $\lim_{N\to\infty} \mathrm{E}^f[\eta^d(X_N^f)|X_0^f = i]$. When d is single class, $\eta^d(i) \equiv \eta^d$ for all $i \in \mathcal{S}$, then $\mathrm{E}^f[\eta^d(X_k^f)|X_0^f = i] \equiv \eta^d$, for all i. The second term

on the right-hand-side of (17) disappears.

Proof We just need to prove (17). By (10) and the assumption, we have

$$\eta^f(i)-\eta^d(i)=$$
$$\lim_{N\to\infty}\frac{1}{N}\sum_{k=0}^{N-1}\mathrm{E}^f\{C^f(X_k^f)|X_0^f=i\}+$$
$$\lim_{N\to\infty}\frac{1}{N}\{\mathrm{E}^f[g^d(X_N^f)|X_0^f=i]-g^d(i)\}-\eta^d(i)=$$
$$\lim_{N\to\infty}\frac{1}{N}\sum_{k=0}^{N-1}\mathrm{E}^f\{C^f(X_k^f)|X_0^f=i\}+$$
$$\lim_{N\to\infty}\frac{1}{N}\sum_{k=0}^{N-1}\mathrm{E}^f\{[P_{\hbar_k}^f g^d](X_k^f)-g^d(X_k^f)|X_0^f=i\}-$$
$$\eta^d(i),$$

in which the assumption (16) plays an important role in the first equality, then the Lemma follows directly from the Poisson equation of g^d. □

4.2 Optimal policy is positive recurrent or multi-chain

Many results follow naturally from Lemma 2. First, we have

Theorem 1 Suppose there is a positive recurrent Markov policy d^* satisfying the Poisson equation

$$A^{d^*}g^{d^*}+C^{d^*}=\eta^{d^*}e, \tag{18}$$

where η^{d^*} is a constant, and

$$\min_{\alpha\in\mathcal{A}(i)}\{C^\alpha(i)+\sum_{j\in\mathcal{S}}P_{i,j}^\alpha g^{d^*}(j)\}=$$
$$C^{d^*}(i)+\sum_{j\in\mathcal{S}}P_{i,j}^{d^*}g^{d^*}(j)=\eta^{d^*}+g^{d^*}(i), \tag{19}$$
$$\mathrm{E}^d[g^{d^*}(X_n^d)|X_0^d=i]<\infty,\ n=1,2,\cdots,$$
$$\lim_{N\to\infty}\frac{1}{N}\mathrm{E}^d[g^{d^*}(X_N^d)|X_0^d=i]=0, \tag{20}$$

for all $i\in\mathcal{S}$ and $d\in\Pi$, then d^* is an optimal policy in Π [5], and the optimality inequality becomes an equality.

Proof Set $r(i):=g^{d^*}(i)$, and $J=\eta^{d^*}$ in the optimality inequality (15), and we get $\eta^{d^*}\leqslant\eta^d(i)$ for all $i\in\mathcal{S}$ and $d\in\Pi$. □

Remark 4 1) In this theorem, d^* cannot be a multi-chain, because J in Lemma 3 has to be a constant. For multi-chain optimal policies, see Theorem 3.

2) d^* is usually positive recurrent, but if it is null recurrent and equation (18) holds for η^{d^*} and a function g^{d^*}, then the theorem may also hold.

3) All the other policies $d\in\Pi$ may be null recurrent, or multi-chain. Condition (19) may not need to hold at some null states which are null-recurrent at all the Markov chains under all other policies.

Theorem 2 Suppose all policies in Π_0 are positive recurrent, and Markov policy $d\in\Pi_0$ is long-run average optimal in Π_0. Then (18) and (19) hold, and it is optimal in Π.

By the performance difference formula (17), we have a set of more general optimality equations. The proof of the following theorem is similar to that in [3] for time nonhomogeneous Markov chains, see also [4] and [8] for finite multi-chains.

Theorem 3 Let $d^*\in\Pi_0$ satisfying the Poisson equation $A^{d^*}g^{d^*}(i)+C^{d^*}(i)=\eta^{d^*}(i)$, and $|\eta^{d^*}(i)|<L<\infty$, for all $i\in\mathcal{S}$. Suppose that for all $\alpha\in\mathcal{A}(i)$, $A_{i,j}^\alpha:=P_{i,j}^\alpha-I$, there exists a constant $M<\infty$ such that

$$|\sum_{j\in\mathcal{S}}A_{i,j}^\alpha g^{d^*}(j)+C^\alpha(i)|\leqslant$$
$$M|\sum_{j\in\mathcal{S}}A_{i,j}^{d^*}g^{d^*}(j)+C^{d^*}(i)|,\ i\in\mathcal{S}, \tag{21}$$

and (20) holds for all $i\in\mathcal{S}, d\in\Pi$. Then d^* is optimal in Π if

a) $$\eta^{d^*}(i)=\min_{\alpha\in\mathcal{A}(i)}\{\sum_{j\in\mathcal{S}}P_{i,j}^\alpha\eta^{d^*}(j)\},\ i\in\mathcal{S}, \tag{22}$$

and there is an $\epsilon>0$ such that for all $i\in\mathcal{S},\alpha\in\mathcal{A}(i)$, if $\eta^{d^*}(i)<\sum_{j\in\mathcal{S}}P_{i,j}^\alpha\eta^{d^*}(j)$ then

$$\sum_{j\in\mathcal{S}}P_{i,j}^\alpha\eta^{d^*}(j)-\eta^{d^*}(i)>\epsilon, \tag{23}$$

and

b) $$\eta^{d^*}(i)+g^{d^*}(i)=$$
$$\min_{\alpha\in\mathcal{A}_0(i)}\{C^\alpha(i)+\sum_{j\in\mathcal{S}}P_{i,j}^\alpha g^{d^*}(j)\},\ i\in\mathcal{S}, \tag{24}$$

where $\mathcal{A}_0(i):=\{\alpha\in\mathcal{A}(i):\sum_{j\in\mathcal{S}}P_{i,j}^\alpha\eta^{d^*}(j)=\eta^{d^*}(i)\}$, $i\in\mathcal{S}$.

Proof Let d be any policy in Π. By (22), we have $\mathrm{E}^d[\eta^{d^*}(X_k^d)|X_0^d=i]\geqslant\eta^{d^*}(i)$, for any k, so we have

$$\lim_{N\to\infty}\frac{1}{N}\sum_{k=0}^{N-1}\mathrm{E}^d[\eta^{d^*}(X_k^d)|X_0^d=i]-\eta^{d^*}(i)\geqslant 0. \tag{25}$$

Next, applying the Dynkin's formula, we have

$$\lim_{K\to\infty}\frac{1}{K}\sum_{k=0}^{K-1}\mathrm{E}^d\{[(P_{\hbar_k}^d\eta^{d^*})(X_k^d)-$$
$$\eta^{d^*}(X_k^d)]|X_0^d=i\}=0.$$

Let I be an indicator function. Under condition (22), the above equation implies

$$\lim_{K\to\infty}\frac{1}{K}\sum_{k=0}^{K-1}\mathrm{E}^d\{[(P_{\hbar_k}^d\eta^{d^*})(X_k^d)-\eta^{d^*}(X_k^d)]$$
$$I[(P_{\hbar_k}^d\eta^{d^*})(X_k^d)-\eta^{d^*}(X_k^d)\geqslant 0]|X_0^d=i\}=0. \tag{26}$$

From $A^{d^*}g^{d^*}(i)+C^{d^*}(i)=\eta^{d^*}(i)$, we group the

[5] (η^{d^*},g^{d^*},d^*) is called a canonical triplet in [7].

terms together

$$\sum_{k=0}^{K-1}[(A^{d^*}g^{d^*})(X_k^d)+C^{d^*}(X_k^d)]=$$
$$\sum_{k=0}^{K-1}\eta^{d^*}(X_k^d)=$$
$$\sum_{k=0}^{K-1}[\eta^{d^*}(X_k^d)]I[(P_{\hbar_k}^d\eta^{d^*})(X_k^d)-\eta^{d^*}(X_k^d)>0]+$$
$$\sum_{k=0}^{K-1}[\eta^{d^*}(X_k^d)]I[(P_{\hbar_k}^d\eta^{d^*})(X_k^d)-\eta^{d^*}(X_k^d)=0]. \tag{27}$$

The third line of (27) is

$$\sum_{k=0}^{K-1}[\frac{\eta^{d^*}(X_k^d)}{(P_{\hbar_k}^d\eta^{d^*})(X_k^d)-\eta^{d^*}(X_k^d)}]$$
$$[(P_{\hbar_k}^d\eta^{d^*})(X_k^d)-\eta^{d^*}(X_k^d)]$$
$$I[(P_{\hbar_k}^d\eta^{d^*})(X_k^d)-\eta^{d^*}(X_k^d)>0].$$

By (23) and the finiteness of $\eta^{d^*}(i)$, the fraction in the above expression is bounded. From (26), we get

$$\lim_{K\to\infty}\frac{1}{K}\sum_{k=0}^{K-1}\mathrm{E}^d\{[(A^{d^*}g^{d^*})(X_k^d)+C^{d^*}(X_k^d)]$$
$$I[(P_{\hbar_k}^d\eta^{d^*})(X_k^d)-\eta^{d^*}(X_k^d)>0]|X_0^d=i\}=0,$$

and therefore, from (27), it holds that

$$\lim_{K\to\infty}\frac{1}{K}\sum_{k=0}^{K-1}\mathrm{E}^d\{(A^{d^*}g^{d^*})(X_k^d)+$$
$$C^{d^*}(X_k^d)|X_0^d=i\}=$$
$$\lim_{K\to\infty}\frac{1}{K}\sum_{k=0}^{K-1}\mathrm{E}^d\{[(A^{d^*}g^{d^*})(X_k^d)+C^{d^*}(X_k^d)]$$
$$I[(P_{\hbar_k}^d\eta^{d^*})(X_k^d)-\eta^{d^*}(X_k^d)\leqslant 0]|X_0^d=i\}. \tag{28}$$

Now, by (21), we have

$$|A_{\hbar_k}^d g^{d^*}(i)+C^d(i)|\leqslant M|A^{d^*}g^{d^*}(i)+C^{d^*}(i)|=$$
$$M|\eta^{d^*}(i)|<ML, \tag{29}$$

for all $i\in\mathcal{S}$ and any history $\hbar_k$. Then similar to (28), we have

$$\lim_{K\to\infty}\frac{1}{K}\sum_{k=0}^{K-1}\mathrm{E}^d\{(A_{\hbar_k}^d g^{d^*})(X_k^d)+$$
$$C^d(X_k^d)|X_0^d=i\}=$$
$$\lim_{K\to\infty}\frac{1}{K}\sum_{k=0}^{K-1}\mathrm{E}^d\{[(A_{\hbar_k}^d g^{d^*})(X_k^d)+C^d(X_k^d)]$$
$$I[(P_{\hbar_k}^d\eta^{d^*})(X_k^d)-\eta^{d^*}(X_k^d)\leqslant 0]|X_0^d=i\}. \tag{30}$$

Furthermore, condition b) is equivalent to

$$(A_{\hbar_k}^d g^{d^*})(i)+C^d(i)\geqslant$$
$$(A^{d^*}g^{d^*})(i)+C^{d^*}(i),\ i\in\mathcal{S},\ d_{\hbar_k}\in\mathcal{A}_0(i). \tag{31}$$

From (31) (28) and (30), we have

$$\lim_{K\to\infty}\frac{1}{K}\mathrm{E}^d\{\sum_{k=0}^{K-1}(A_{\hbar_k}^d g^{d^*})(X_k^d)+$$
$$C^d(X_k^d)|X_0^d=i\}=$$
$$\lim_{K\to\infty}\frac{1}{K}\mathrm{E}^d\{\sum_{k=0}^{K-1}[(A_{\hbar_k}^d g^{d^*})(X_k^d)+C^d(X_k^d)]$$
$$I[(P_{\hbar_k}^d\eta^{d^*})(X_k^d)-\eta^{d^*}(X_k^d)\leqslant 0]|X_0^d=i\}=$$
$$\lim_{K\to\infty}\frac{1}{K}\mathrm{E}^d\{\sum_{k=0}^{K-1}[(A_{\hbar_k}^d g^{d^*})(X_k^d)+C^d(X_k^d)]$$
$$I[(P_{\hbar_k}^d\eta^{d^*})(X_k^d)-\eta^{d^*}(X_k^d)=0]|X_0^d=i\}\geqslant$$
$$\lim_{K\to\infty}\frac{1}{K}\mathrm{E}^d\{\sum_{k=0}^{K-1}[(A^{d^*}g^{d^*})(X_k^d)+C^{d^*}(X_k^d)]$$
$$I[(P_{\hbar_k}^d\eta^{d^*})(X_k^d)-\eta^{d^*}(X_k^d)\leqslant 0]|X_0^d=i\}=$$
$$\lim_{K\to\infty}\frac{1}{K}\mathrm{E}^d\{\sum_{k=0}^{K-1}[(A^{d^*}g^{d^*})(X_k^d)+$$
$$C^{d^*}(X_k^d)]|X_0^d=i\}. \tag{32}$$

Next, setting $f:=d$ and $d:=d^*$ in the performance difference formula (17), we get

$$\eta^d(i)-\eta^{d^*}(i)=$$
$$\lim_{N\to\infty}\frac{1}{N}\sum_{k=0}^{N-1}\mathrm{E}^d\{\{[P_{\hbar_k}^d g^{d^*}](X_k^d)+C^d(X_k^d)\}-$$
$$\{[P^{d^*}g^{d^*}](X_k^d)+C^{d^*}(X_k^d)\}|X_0^d=i\}+$$
$$\{\lim_{N\to\infty}\frac{1}{N}\sum_{k=0}^{N-1}\mathrm{E}^d[\eta^{d^*}(X_k^d)|X_0^d=i]-\eta^{d^*}(i)\}. \tag{33}$$

From this equation, (25) and (32), we conclude that $\eta^d(i)\geqslant\eta^{d^*}(i)$, $i\in\mathcal{S}$, or $\eta^d\geqslant\eta^{d^*}$, for all $d\in\Pi$. That is, d^* is an optimal policy. □

Remark 5 1) The restriction of (24) to $\mathcal{A}_0$ is very important; otherwise, the inequality in (32) may be false and conditions (22) and (24) may not have a solution, see also Example 9.1.1 in [11].

2) The technical conditions (21) and (23) indicate that some kind of uniformity is required for countable-state problems. They are not very restrictive and can be replaced by other similar conditions.

4.3 Under-selectivity

Under-selectivity refers to the property that the long-run average of a stochastic chain does not depend on its value in any finite period, or at any "non-frequently visited" periods, or states, defined below.

Definition 1 a) A subsequence of $k=0,1,\cdots$, denoted by $k_0,k_1,\cdots$, is called a non-frequently visited sequence, if $\lim_{n\to\infty}\frac{n}{k_n}=0$.

b) A subset of the state space, $\mathcal{S}_0\subset\mathcal{S}$, is called a non-frequently visited set of a Markov chain $\boldsymbol{X}$, if

$$\lim_{K\to\infty}\frac{1}{K}\sum_{k=0}^{K-1}I(X_k\in\mathcal{S}_0)=0,\ \text{w.p.1}, \tag{34}$$

with I being an indicator function.

It is clear that non-frequently visited sets include transient states in multi-chains and finite sets of null recurrent states.

Theorem 4 Theorem 3 holds even if Condition

(24) does not hold at any non-frequently visited set of states of a policy.

Proof By (29), $|(\eta^{d^*}(i)+g^{d^*}(i))-(C^{\alpha}(i)+\sum\limits_{j\in\mathcal{S}}P^{\alpha}_{i,j}g^{d^*}(j))|$, $\alpha\in\mathcal{A}_0(i)$, $i\in\mathcal{S}$, are bounded. Suppose (24) does not hold on a non-frequently visited set $\mathcal{S}_0$ of states of a policy d_0. Then by (34), we have

$$\lim_{K\to\infty}\frac{1}{K}\sum_{k=0}^{K-1}I(X_k^{d_0}\in\mathcal{S}_0)$$
$$[\mathrm{E}^d\{\{[P^d_{\hbar_k}g^{d^*}](X^d_k)+C^d(X^d_k)\}-$$
$$\{[P^{d^*}g^{d^*}](X^d_k)+C^{d^*}(X^d_k)\}|X^d_0=i\}]=0,\ \text{w.p.1}.$$

This cannot change the relation $\eta^d\geqslant\eta^{d^*}$ in (33). □

4.4 Null recurrent Markov policies

When the optimal policy is null recurrent, the situation is a bit complicated. The Poisson equation (3) and the potential function $g(i)$ may not exist, and it is difficult to find the function $r(i)$, $i\in\mathcal{S}$, in (13), except for very simple cases, e.g. in Examples 1 and 2, $r(i)\equiv 0$, for all $i\in\mathcal{S}$.

To further understand the problem, consider a null recurrent policy f, and let d be any positive recurrent Markov policy with the potential function $g^d(i)$, $i\in\mathcal{S}$. Assume that (13) (with the $\geqslant$ sign) holds for $J=\eta^d$, $r(i)=g^d(i)$, $\alpha=f(i)$, $C^f(i)=C^{\alpha}(i)$, i.e., $\eta^d+g^d(i)\geqslant C^{\alpha}(i)+\sum\limits_{j\in\mathcal{S}}P^{\alpha}_{i,j}g^d(j)$. Then $\eta^d\geqslant\eta^f$, i.e., $\eta^f=\min\{\eta^f,\eta^d\}$. By the Poisson equation for d, we have $C^d(i)+\sum\limits_{j\in\mathcal{S}}P^d_{i,j}g^d(j)\geqslant C^f(i)+\sum\limits_{j\in\mathcal{S}}P^f_{i,j}g^d(j)$. Therefore, we have

$$\eta^d+g^d(i)\geqslant\min\{xC^f(i)+\sum_{j\in\mathcal{S}}P^f_{i,j}g^d(j),\ C^d(i)+\sum_{j\in\mathcal{S}}P^d_{i,j}g^d(j)\}. \quad (35)$$

This is the "optimality inequality" among these two policies. As explained in Examples 1 and 2, this inequality is due to the null recurrency of i and the underselectivity of average optimality. In fact, both directions $\leqslant$ and $\geqslant$ are possible in (35). When d is also null recurrent, the situation is more complicated because there is no Poisson equation.

The next specially designed examples illustrates the application of Theorem 1 to a null recurrent policy. We need use the zeta function $\zeta(s)=\sum\limits_{k=1}^{\infty}\frac{1}{k^s}$. It is known that $\zeta(1)=\infty$, $\zeta(\frac{3}{2})\approx 2.6124$, $\zeta(2)=\frac{\pi^2}{6}\approx 1.6449$, $\zeta(3)\approx 1.2021$.

Example 4 The structure of the Markov chain is the same as Example 1. The state space is $\mathcal{S}=\{0,1,2,\cdots\}$. At state $i\geqslant 1$, there is a null action with $P_{i,i-1}=1$. At state 0, the transition probabilities are given by $P_{0,i}>0$, $i\geqslant 1$. We consider two policies f and d:

1) For policy f, the transition probabilities at state 0 is $P^f_{0,i}=p_i$, $i\geqslant 1$, with $\sum\limits_{i=1}^{\infty}p_i=1$, $\sum\limits_{i=1}^{\infty}ip_i=\infty$, $\sum\limits_{i=1}^{\infty}\sqrt{i}p_i<\infty$. For example, we may take $p_i=\frac{1}{\zeta(2)}\frac{1}{i^2}$. The cost function is set $C^f(i)\equiv 1$, for all $i\in\mathcal{S}$.

As shown in Example 1, the Markov chain is null recurrent, with long-run average $\eta^f=1$.

2) For policy d, the transition probabilities at state 0 is set to be $P^d_{0,i}=q_i$, $i\geqslant 1$, with $\sum\limits_{i=1}^{\infty}q_i=1$, $\sum\limits_{i=1}^{\infty}iq_i<\infty$, $\sum\limits_{i=1}^{\infty}\sqrt{i}q_i<\infty$. For example, we may take $q_i=\frac{1}{\zeta(2.5)}\frac{1}{i^{2.5}}$. So the first passage time $\tau(0,0)$ is finite and the Markov chain under d is positive recurrent.

We define the cost function for d for $i>0$ by

$$C^d(i)=\begin{cases}1, & \text{if } i=k^2 \text{ for some integer } k\geqslant 1,\\ 0, & \text{otherwise},\end{cases}$$

and choose a special value for $C^d(0)$ so that $\eta^d=0$. Let π_i be the steady-state probability at state i, $i=0,1,\cdots$. Thus, we choose

$$C^d(0)=-\frac{\sum\limits_{i=1}^{\infty}\pi_iC^d(i)}{\pi_0}. \quad (36)$$

By the structure of the Markov chain, every time it visits state 0, it has to visit state i, $\sum\limits_{k=i}^{\infty}q_k$ times ($i=1,2,\cdots$). Therefore, $\pi_i=(\sum\limits_{k=i}^{\infty}q_k)\pi_0$, and thus,

$$\pi_0=\frac{1}{1+\sum\limits_{i=1}^{\infty}\sum\limits_{k=i}^{\infty}q_k}=\frac{1}{1+\sum\limits_{k=1}^{\infty}kq_k}>0.$$

By (36), we have

$$C^d(0)=-\sum_{i=1}^{\infty}(\sum_{k=i}^{\infty}q_k)C^d(i). \quad (37)$$

Next, we choose $z=0$ as the reference state, with $g^d(0)=0$. By (2), the potential function at $i>0$ is

$$g^d(i)=\mathrm{E}^d\{\sum_{k=0}^{i-1}[C^d(X^d_k)]|X^d_0=i\}=\sum_{k=1}^{i}C^d(k)=\lfloor\sqrt{i}\rfloor, \quad (38)$$

where $\lfloor x\rfloor$ denotes the largest integer less or equal x.

Now, we are ready to check the Poisson equation (3). At $i=k^2$ for some integer k,

$$g^d(i)=k,\ g^d(i-1)=k-1,$$
$$P^d_{i,i-1}=1,\ C^d(i)=1.$$

(3) is $g^d(i-1)-g^d(i)+C^d(i)=\eta^d$, or $(k-1)-$

$k+1=0$, which indeed holds. At $i=k^2+1$, it is

$$g^d(k^2)-g^d(k^2+1)+C^d(k^2+1)=\eta^d,$$

or $k-k+0=0$. At $i>0$ but not one of the above two cases,

$$g^d(i)=\lfloor\sqrt{i}\rfloor=\lfloor\sqrt{i-1}\rfloor=g^d(i-1),$$

so it is $\lfloor\sqrt{i-1}\rfloor-\lfloor\sqrt{i}\rfloor+0=0$. In all these cases for $i>0$, (3) holds. Finally, we verify that (3) holds at $i=0$. Indeed, by (38) and (37), it holds

$$A^dg^d(0)=\sum_{k=1}^{\infty}q_kg^d(k)-g^d(0)=$$

$$\sum_{k=1}^{\infty}[q_k\sum_{i=1}^{k}C^d(i)]=\sum_{i=1}^{\infty}[\sum_{k=i}^{\infty}q_kC^d(i)]=-C^d(0).$$

We have $\eta^d=0<1=\eta^f$, so d is optimal among $\{d,f\}$. Let us verify the optimality equation (19). First, we have

$$H_0:=\mathrm{E}^f[g^d(X_1^f)|X_0^f=0]=\sum_{i=1}^{\infty}p_ig^d(i)=$$

$$\sum_{i=1}^{\infty}\frac{\lfloor\sqrt{i}\rfloor}{\zeta(2)i^2}<\sum_{i=1}^{\infty}\frac{\sqrt{i}}{i^2}=\sum_{i=1}^{\infty}(\frac{1}{i})^{\frac{3}{2}}<\infty. \tag{39}$$

Thus,

$$(A^fg^d)(0)+C^f(0)=H_0+1>0=$$
$$(A^dg^d)(0)+C^d(0).$$

Also, it is easy to check that

$$(P^fg^d)(i)+C^f(i)=g^d(i-1)+C^f(i)\geqslant$$
$$g^d(i-1)+C^d(i)=(P^dg^d)(i)+C^d(i),\ i>0.$$

That is, the optimality equation (19) holds at all states $i\in\mathcal{S}$. So we have $\eta^f>\eta^d$.

It is a bit tedious to check the condition (12). We briefly discuss it as follows. First, we define a vector denoted as W_n, $n=0,1,2,\cdots$, whose ith component is

$$\mathrm{E}^f[g^d(X_n^f)|X_0^f=i],\ i=0,1,\cdots.$$

Thus,

$$W_0=(g^d(0)\ \cdots\ g^d(i)\ \cdots)^{\mathrm{T}}=(0\ 1\ \cdots\ \lfloor\sqrt{i}\rfloor\ \cdots)^{\mathrm{T}}.$$

By (39) and the structure of $P_{i,j}^f$, we have $W_1=(H_0\ 0\ 1\ \cdots\ \lfloor\sqrt{i}\rfloor\ \cdots)^{\mathrm{T}}=(H_0\ W_0^{\mathrm{T}})^{\mathrm{T}}$. In W_1, the vector W_0 shifts by one step towards right. Then

$$W_2=(H_1,W_1^{\mathrm{T}})^{\mathrm{T}}=(H_1,H_0,W_0^{\mathrm{T}})^{\mathrm{T}},$$

where

$$H_1=\sum_{i=1}^{\infty}\frac{\lfloor\sqrt{i}\rfloor}{\zeta(2)(i+1)^2}<\sum_{i=1}^{\infty}\frac{\sqrt{i}}{(i+1)^2},$$

and we have $W_3=(H_2\ H_1\ H_0\ W_0^{\mathrm{T}})^{\mathrm{T}}$, where

$$H_2<H_1+H_0+\sum_{i=1}^{\infty}\frac{\sqrt{i}}{(i+2)^2}<\sum_{n=0}^{2}\sum_{i=1}^{\infty}\frac{\sqrt{i}}{(i+n)^2}.$$

In general, we have

$$W_N=(H_{N-1}\ H_{N-2}\ \cdots\ H_1\ H_0\ W_0^{\mathrm{T}})^{\mathrm{T}},$$

where $H_N<\sum\limits_{n=0}^{N}\sum\limits_{i=1}^{\infty}\dfrac{\sqrt{i}}{(i+n)^2}$. Finally, by Stolz theorem, we have

$$\lim_{N\to\infty}\frac{H_N}{N}\leqslant\lim_{N\to\infty}\frac{\sum\limits_{n=0}^{N}\sum\limits_{i=1}^{\infty}\dfrac{\sqrt{i}}{(i+n)^2}}{N}=$$
$$\lim_{N\to\infty}\sum_{i=1}^{\infty}\frac{\sqrt{i}}{(i+N)^2}=0.$$

Condition (12) is thus proved.

5 Conclusion

We summarize this paper with the following observations.

1) The optimality inequality (13) is a sufficient condition for average optimal. The strict inequality may hold at null recurrent states. When the optimal policy is positive recurrent, the inequality becomes an equality.

2) A null recurrent Markov chain visits any state with a probability of zero, so the cost function at such a state can be changed without changing the value of long-run average. Thus, any optimality equality or inequality involving the cost may not need to hold at such a state. In other words, the inequality may be in either direction at a null recurrent state for the optimal policy (see Examples 1 and 2). This is purely the consequence of null recurrency and under-selectivity.

3) The existence of average optimal policies of countable MDPs is mainly derived by the Dynkin's formula from a view of performance difference. Example 4 shows the application of the main results, which makes a supplement to the existing literature work.

References:

[1] ALTMAN E. *Constrained Markov Decision Processes*. Boca Raton: Chapman & Hall/CRC Press, 1999.

[2] BERTSEKAS D P. *Dynamic Programming and Optimal Control: Volumes I and II*. Belmont, Massachusetts: Athena Scientific, 2005.

[3] CAO X R. *Foundations of Average-Cost Nonhomogeneous Controlled Markov Chains*. New York: Springer, 2020.

[4] CAO X R. *Stochastic Learning and Optimization – A Sensitivity-Based Approach*. New York: Springer, 2007.

[5] FEINBERG E A, SHWARTZ A (eds.). *Handbook of Markov Decision Processes: Methods and Application*. Boston: Kluwer Academic Publishers, 2002.

[6] GUO X P, HERNÁNDEZ-LERMA O. *Continuous-Time Markov Decision Processes*. New York: Springer, 2009.

[7] HERNÁNDEZ-LERMA O, LASSERRE J B. *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. New York: Springer, 1996.

[8] LAURENCE A B. Markov decision processes: discrete stochastic dynamic programming. *Technometrics*, 1995, 37(3): 353 – 353.

[9] ROSS S. *Introduction to Stochastic Dynamic Programming*. New York: Academic Press, 1983.

[10] SENNOTT L I. *Stochastic Dynamic Programming and the Control of Queueing Systems*. New Jersey: Wiley, 1999.

[11] SENNOTT L I. Average cost optimal stationary policies in infinite state Markov decision processes with unbounded costs. *Operations Research*, 1989, 37: 626 – 633.

[12] XIA L, GUO X P, CAO X R. On the existence of optimal stationary policies for average Markov decision processes with countable states. *Manuscript*, 2020.

[13] FLEMING W H, SONER H M. *Controlled Markov Processes and Viscosity Solutions*. New York: Springer, 2006.

[14] HOPP W J, BEAN J C, SMITH R L. A new optimality criterion for nonhomogeneous Markov decision processes. *Operations Research*, 1987, 35(6): 875 – 883.

[15] CAO X R. Optimization of average rewards of time nonhomogeneous Markov chains. *IEEE Transactions on Automatic Control*, 2015, 60: 1841 – 1856.

[16] NI Y H, FANG H T. Policy iteration algorithm for singular controlled diffusion processes. *SIAM Journal on Control and Optimization*, 2013, 51: 3844 – 3862.

[17] CAO X R, CHEN H F. Potentials, Perturbation realization, and sensitivity analysis of Markov processes. *IEEE Transactions on Automatic Control*, 1997, 42: 1382 – 1393.

作者简介:

张俊玉 副教授, 博士生导师, 1999年获北京大学数学学院概率统计系概率论与数理统计专业学士学位, 2002年获中国科学院数学与系统科学研究院硕士学位, 2006年获香港科技大学电机与电子工程学博士学位, 目前研究方向为马尔可夫决策过程、随机优化、随机博弈和离散事件动态系统, E-mail: mcszhjy@mail.sysu.edu.cn;

吴怡婷 硕士研究生, 2020年获中山大学数学与应用数学专业学士学位, 目前研究方向为马尔可夫决策过程、随机优化和随机博弈, E-mail: wuyt35@mail2.sysu.edu.cn;

夏 俐 教授, 博士生导师, 分别于2002年和2007年获清华大学控制理论学士和博士学位, 目前研究方向为随机学习与优化方法研究、马尔可夫决策过程、强化学习、排队理论以及在能源系统、金融科技等方面的应用研究, E-mail: xiali5@sysu.edu.cn;

曹希仁 教授, 博士生导师, 1984年获美国哈佛大学应用数学博士学位, 他是美国数字设备公司的咨询工程师, 哈佛大学的研究员, 以及香港科技大学的教授和讲座教授, 现为上海交通大学客座教授, 曹希仁博士是IFAC会员和IEEE终身会员, 目前研究方向为随机控制、金融工程、随机学习与优化、离散事件动态系统, E-mail: eecao@ust.hk.