Turnpikes in Finite Markov Decision Processes and Random Walk

Alexey B. Piunovskiy

June 14, 2022

University of Liverpool
Dept. of Mathematical Sciences, Liverpool L69 7ZL, UK
piunov@liverpool.ac.uk

Abstract

In this paper we revise the theory of turnpikes in discounted Markov decision processes, prove the Turnpike Theorem for the undiscounted model and apply the results to the specific random walk.

Keywords: turnpike, Markov decision process, discounted reward, average reward, random walk, stochastic knapsack problem
AMS 2020 subject classification: Primary 90C40, Secondary 90C39, 60G50

1 Introduction

It looks, the term ‘Turnpike’ first appeared in [4]:

> ”... *we have found a real normative significance for steady growth ... if we are planning long-run growth, no matter where we start, and where we desire to end up, it will pay in the intermediate stages to get into a growth phase of this kind. It is exactly like a turnpike paralleled by a network of minor roads.*”

That book, like many other early works on turnpikes, was about economical models.

In the recent monograph [26], the modern theory of deterministic turnpikes was also developed in the strict connection with economics. The basic mathematical model, which can give the idea of turnpikes, looks like

$$\sum_{i=0}^{T-1} v(x_i, x_{i+1}) \to \max \text{ over } \{x_i\}_{i=0}^{T} \in \Omega,$$

where x_i are elements of some state space and v is a real-valued function. The turnpike trajectory $\bar{X} := (\bar{x}, \bar{x}, \ldots, \bar{x})$, which is optimal in the long middle part of the planning horizon $1, 2, \ldots, T$, comes from the maximization of $v(x, x)$. Of course, it is assumed that $\bar{X} \in \Omega$ and some other conditions are satisfied. More about such models in [24, Ch.8].

The continuous-time analogue is the variational problem:

$$\text{minimize } \int_0^T f(x(t), x'(t))\, dt \text{ over}$$
$$\text{absolutely continuous functions } x:\ [0,T] \to \mathbb{R}^n \text{ with } x(0) = x,\ x(T) = y.$$

Under appropriate conditions, the turnpike trajectory is just $x(t) \equiv X_f = \arg\min\{f(x,0)\}$ [24]. Again, this simple function is optimal in the long middle part of the interval $[0,T]$.

Monograph [25] is devoted to deterministic optimal control problems, with the state and action spaces $\mathbb{R}^n$ and $\mathbb{R}^m$, of the shape

$$\begin{aligned} &\int_0^T f(t,x,u)\ dt \longrightarrow \inf \qquad (1)\\ &\text{subject to } x' = G(t,x,u),\ x(0)=x,\ x(T)=y. \end{aligned}$$

Without going in details, under appropriate conditions, Theorem 2.3 of [25] says that there exists a turnpike $X_f:\ \mathbb{R}_+ \to \mathbb{R}^n$ satisfying the following property. For each $\varepsilon > 0$ there exist $\delta \in (0,\varepsilon)$ and $\Delta > 0$ such that for all $T \geq 2\Delta$ $\|x(t) - X_f(t)\| \leq \varepsilon$ for all $t \in [\Delta, T-\Delta]$, for each trajectory $x:\ [0,T] \to \mathbb{R}^n$, δ-optimal in problem (1) with arbitrary enough $x(0)$ and $x(T)$, and an arbitrary cost rate g close enough to f. Moreover, in the infinite-horizon case $T=\infty$, $\lim_{t\to\infty}\|x(t)-X_f(t)\| = 0$ [25, Prop.3.2]. Similar optimal control problems with linear function G were investigated in [24, Ch.6,7], where stronger turnpike properties were established.

The theory of turnpikes was also developed for piece-wise deterministic processes [7] and for continuous-time Markov decision processes (MDPs) [22] with applications to manufacturing. The versions for controlled diffusions and more general random processes with financial applications can be found in [6]. Thus, the turnpike property is a general phenomenon which holds for large classes of variational and optimal control problems. Moreover, it was shown in [24, 25] that the turnpike is not necessarily unique.

The further material of the present article is devoted to discrete-time MDPs. In this framework, the turnpike property means the following.

> If the time horizon is large enough, then the optimal control on the first several steps is also optimal for the infinite-horizon version of the model.

Such property is well known for MDPs with the total expected discounted rewards [5, 8, 9, 15, 21, 23]. The versions with the exponential utility function were considered in [3]. Instead of discounting, one can consider the random time horizon [10] or other special cases [11]: see Remark 1. The state and action spaces usually were finite [3, 5, 9, 11, 15, 21, 23].

Note that, differently from the model (1), here the turnpike control is optimal on the initial interval $\{1,2,\dots,T-\Delta\}$ rather than on $\{\Delta,\Delta+1,\dots,T-\Delta\}$.This is because here the turnpike property is formulated for the feedback control rather than for trajectories.

In the present article, we consider the finite model with the total expected (undiscounted) reward associated with the long-run average reward on the infinite time horizon. The most relevant references are [8, 14, 17]. Note that the model was not finite in [8, 14], but in the case of finite MDP the results of the aforementioned articles are similar to those presented below in Sections 3-5: see Remarks 2 and 3.

The novelty of the present article is in the following:

- we consider not only the standard MDP, but more general finite models;
- we show that, for the discounted model with the discount factor close to 1, there may exist two turnpikes; such phenomenon was not observed before;
- we apply the developed theory to the random walk which is of its own interest.

In Section 2, we remind the known results on the standard turnpikes in discounted MDPs. In Section 3, we develop the turnpike theory for the undiscounted model. This is the main part of the article. The proofs of the main results are presented in Sections 4 and 5. In Section 6 we return back to the discounted model and show that, if the discount factor is close to 1, then one can observe two turnpikes. In Section 7 we consider an example of controlled random walk which exhibits the turnpike property. In the summary, we enlist what has been done and fix important open problems. The proofs of auxiliary statement are postponed to Appendix.

All over the current paper, the state and action spaces $\mathbf{S}$ and $\mathbf{A}$ are finite. For the future needs, it is convenient to accept that $\mathbf{S} = \{0, 1, \dots, M-1\}$, $M \ge 1$. Below, the following notations are in use: $e := (1, 1, \dots, 1)^T \in \mathbb{R}^M$, and I is the identity $M \times M$ matrix. We say that a stochastic matrix P is irreducible or aperiodic, if the corresponding Markov chain is so. W_s is the s-th component of the vector $W \in \mathbb{R}^M$, and $P_{s,\cdot}$ is the s-th row of the square $M \times M$ matrix P. Each function $W : \mathbf{S} \to \mathbb{R}$ is identified with the vector $W \in \mathbb{R}^M$, so the both notations $W(s) = W_s$ are in use. All vectors are columns; all inequalities for vectors and matrices are component-wise. $\mathbf{0}$ is the zero vector in $\mathbb{R}^M$.

2 Standard Turnpikes in MDP

To introduce the idea of a turnpike for the discrete-time MDP, let us consider the discounted model $\langle \mathbf{S}, \mathbf{A}, P, r \rangle$ with the infinite horizon, the finite state and action spaces $\mathbf{S}$ and $\mathbf{A}$, the one-step reward $r_s(a)$, and the transition probability $P_{s,j}(a)$. This means that every time action $a \in \mathbf{A}$ is applied at state $s \in \mathbf{S}$, the reward equals $r_s(a)$ and the new state of the process is realized according to the distribution $P_{s,\cdot}(a)$. The optimality equation looks like

$$W^\infty(s) = \max_{a \in \mathbf{A}} \left\{ r_s(a) + \beta \sum_{j \in \mathbf{S}} P_{s,j}(a) W^\infty(j) \right\}, \quad s \in \mathbf{S}, \tag{2}$$

where $\beta \in (0,1)$ is the discount factor. It is well known that this equation has a unique bounded solution which can be constructed by the successive approximations called 'value iteration':

$$\begin{aligned} W^0 &= W^{term} \ \text{ is an arbitrarily fixed real-valued function on } \mathbf{S}; \\ W^{k+1}(s) &= U^\beta \circ W^k(s) := \max_{a \in \mathbf{A}} \left\{ r_s(a) + \beta \sum_{j \in \mathbf{S}} P_{s,j}(a) W^k(j) \right\}, \quad s \in \mathbf{S}, \\ & \quad k = 0, 1, 2, \dots; \end{aligned} \tag{3}$$

$\lim_{k\to\infty} \max_{s \in \mathbf{S}} |W^k(s) - W^\infty(s)| = 0$. We use the notations W^0 and W^{term} for the same function because, on one hand, it is the starting point for iterations (3), and on the other hand that function describes the terminal reward in the finite horizon model. (See below.) The operator U^β is a contraction in the space of real-valued functions W on $\mathbf{S}$ (which are certainly bounded) with the uniform norm $\|W\| := \max_{s \in \mathbf{S}} |W(s)|$. A control strategy is (uniformly) optimal if and only if, in each state $s \in \mathbf{S}$, the decision maker applies action(s) from the set

$$D_*(s) := \left\{ a \in \mathbf{A} : \quad r_s(a) + \beta \sum_{j \in \mathbf{S}} P_{s,j}(a) W^\infty(j) = W^\infty(s) \right\} \neq \emptyset.$$

See [18, §1.2.2] or [21, Ch.6]. Usually, one puts $W^0(s) \equiv 0$, but it is convenient to allow W^0 to be arbitrary. The function W^∞ certainly does not depend on W^0.

Since the spaces $\mathbf{S}$ and $\mathbf{A}$ are finite, the value

$$\Delta := \min_{s \in \mathbf{S}} \min_{a \in \mathbf{A} \setminus D_*(s)} \left\{ W^\infty(s) - \left[r_s(a) + \beta \sum_{j \in \mathbf{S}} P_{s,j}(a) W^\infty(j) \right] \right\}$$

is strictly positive. (Here, as usual, $\min_{a\in\emptyset} G(a) := +\infty$.)

If $\Delta = +\infty$, then all $a \in \mathbf{A}$ provide the maximum in (2) and $D_*(s) = \mathbf{A}$ for all $s \in \mathbf{S}$, and obviously, for all $k \geq 0$,

$$\arg\max_{a\in\mathbf{A}} \left\{ r_s(a) + \beta \sum_{j\in\mathbf{S}} P_{s,j}(a) W^k(j) \right\} \subset D_*(s) \quad \text{for all } s \in \mathbf{S}. \tag{4}$$

Suppose $\Delta < \infty$. Then, for each $s \in \mathbf{S}$ and each $a \in \mathbf{A} \setminus D_*(s)$,

$$\left[r_s(a) + \beta \sum_{j\in\mathbf{S}} P_{s,j}(a) W^\infty(j) \right] \leq W^\infty(s) - \Delta.$$

Let us choose $0 < \varepsilon < \frac{\Delta}{2}$ and fix K such that

$$\max_{j\in\mathbf{S}} |W^k(j) - W^\infty(j)| < \varepsilon \quad \text{for all } k \geq K. \tag{5}$$

Now, for each $k \geq K$, for each $s \in \mathbf{S}$, if $a \notin D_*(s)$ provides the maximum in (3), then

$$\begin{aligned} W^{k+1}(s) &= r_s(a) + \beta \sum_{j\in\mathbf{S}} P_{s,j}(a) W^k(j) \leq r_s(a) + \beta \sum_{j\in\mathbf{S}} P_{s,j}(a) W^\infty(j) + \varepsilon \\ &\leq W^\infty(s) - \Delta + \varepsilon < W^{k+1}(s) + \varepsilon - \Delta + \varepsilon < W^{k+1}(s). \end{aligned}$$

The obtained contradiction shows that, for all $k \geq K$, we have inclusion (4).

We have established the classical (discounted) **Turnpike Theorem:**

There exists K such that, for all $k \geq K$, the inclusion (4) holds.

The minimal value of such K is called the 'turnpike integer' and denoted as K^*.

This theorem appeared in [23]; see also [21, Thm.6.8.1], [5] and [15], where the upper bound of K^* was calculated and the dependence of K^* on $\beta \in (0,1)$ was discussed.

Consider the discounted MDP with the finite horizon and the fixed terminal reward W^{term} which is independent of the horizon T:

$$W_T^\pi(\hat{s}) := \mathbb{E}_{\hat{s}}^\pi \left[\sum_{t=1}^{T} \beta^{t-1} r_{S(t-1)}(A(t)) + \beta^T W^{term}(S(T)) \right] \to \sup_\pi .$$

Here $\hat{s} \in \mathbf{S}$ is the initial state $S(0)$, π is the control strategy (policy), $\mathbb{E}_{\hat{s}}^\pi$ is the mathematical expectation with respect to the strategical measure $\mathbb{P}_{\hat{s}}^\pi$ on $\Omega := \mathbf{S} \times (\mathbf{A} \times \mathbf{S})^n$, and $S(t)$, $A(t)$ are the (random) controlled and action processes, i.e., the projection functions on Ω. The detailed rigorous constructions can be found in [18, 21].

According to the dynamic programming principle, the Bellman function

$$W^T(s) := \sup_\pi W_T^\pi(s), \quad T = 0, 1, \ldots$$

satisfies equalities (3) and, for fixed $T \geq 1$, the optimal values of $A(1), A(2), \ldots, A(T)$ after observing $S(0), S(1), \ldots, S(T-1)$ are those which provide the maximum in (3) at $k = T-1, T-2, \ldots, 0$, $s = S(0), S(1), \ldots, S(T-1)$ correspondingly. According to the Turnpike Theorem, if $T > K^*$, then on the first steps $t = 1, 2, \ldots, T - K^*$, having observed $S(t-1)$, one has to choose the actions $A(t)$ from the set $D_*(S(t-1))$. This observation is useful if $D_*(s)$ is a singleton for all $s \in \mathbf{S}$: for large time horizon T, the optimal feedback control on the first steps $t = 1, 2, \ldots, T - K^*$ is the same, independent of the terminal reward W^{term} and is represented by the mapping D_*. This is the so called 'turnpike'. In words, the Turnpike Theorem reads:

$$\text{If the time horizon is large, on the first steps one has} \\ \text{to control the process as if the horizon is infinite.} \tag{6}$$

Close to the end, on the steps $T - K^* + 1, T - K^* + 2, \ldots, T$, the control process $A(t)$ is time-dependent, changing as time t goes on, because one has to take into account the finite remaining horizon and the terminal reward W^{term}. In case the set $D^*(\hat{s}) = \{a_1, a_2, \ldots\}$ is not a singleton for some $\hat{s} \in \mathbf{S}$, it can happen that, for each $T > K^*$, either a_1 or a_2 is not optimal at the first step in the finite horizon MDP with the initial state $\hat{s}$ [23].

Remark 1 *It is well known that the discounting can be interpreted as the geometrical random time horizon: see [18, §1.2.2]. Turnpike theorems for models with other random time horizons were studied in [10]. In [11], the time horizon consists of (long enough) cycles, and during each cycle the optimal feedback strategy is of the turnpike shape.*

3 Description of the Model and Main Results

In case $\beta = 1$, the turnpike theory is more problematic because the operator U^1 is not a contraction in the uniform norm and one cannot in general expect the convergence of the sequence $\{W^k\}_{k=0}^{\infty}$ to a bounded function. Below, we consider the slightly more general iterations than (3). Namely, let $\mathbf{D}$ be the finite space of 'decisions', each decision leading to the (column) vector of rewards $\mathcal{R}(d) \in \mathbb{R}^M$ and to the stochastic $M \times M$ matrix $\mathcal{Q}(d)$. The triplet $\langle \mathbf{D}, \mathcal{R}, \mathcal{Q} \rangle$ will be called 'the model'. Everywhere further we assume that the introduced objects satisfy the following requirement.

Assumption 1 *For every fixed $\beta \in [0, 1]$, for each function $W : \mathbf{S} \to \mathbb{R}$, there exists $\hat{d} \in \mathbf{D}$ providing the following component-wise maximum:*

$$\mathcal{R}(d) + \beta \mathcal{Q}(d) W \to \max_{d \in \mathbf{D}}.$$

The value iterations (3) are generalized to

$$\begin{aligned} W^0 &= W^{term} \text{ is an arbitrarily fixed real-valued function on } \mathbf{S}; \\ W^{k+1} &= U^\beta \circ W^k := \max_{d \in \mathbf{D}} \left\{ \mathcal{R}(d) + \beta \mathcal{Q}(d) W^k \right\}, \\ & \quad k = 0, 1, 2, \ldots, \end{aligned} \tag{7}$$

and the undiscounted case corresponds to $\beta = 1$ in which case we omit the β index: $U := U^1$. In Section 6, the functions as in (7) are denoted $W^{\beta k}$ if $\beta \in (0, 1)$.

Let us describe special cases, where, like previously, $\mathbf{A}$ is a finite action space, for each $a \in \mathbf{A}$, $P(a) = [P_{s,j}(a)]_{(s,j) \in \mathbf{S}^2}$ is a stochastic $M \times M$ matrix and $r(a) = [r_s(a)]_{s \in \mathbf{S}} \in \mathbb{R}^M$ is a column vector. To put it different, let $\langle \mathbf{S}, \mathbf{A}, P, r \rangle$ be an MDP and accept that $\mathbf{D}$ is the set of all mappings $d : \mathbf{S} \to \mathbf{A}$.

(i) Standard model:

$$\begin{aligned} \mathcal{R}(d) &= [r_s(d(s))]_{s \in \mathbf{S}}; \\ \mathcal{Q}(d) &= [P_{s,j}(d(s))]_{(s,j) \in \mathbf{S}^2}. \end{aligned}$$

In this case iterations (7) represent the standard value iteration algorithm for MDP, the same as (3). Assumption 1 is obviously satisfied: the mapping $\hat{d} \in \mathbf{D}$ is built separately for all $s \in \mathbf{S}$.

(ii) More general case:

$$\begin{aligned}
\mathcal{R}_0(d) &= r_0(d(0)); \\
\mathcal{R}_{l+1}(d) &= r_{l+1}(d(l+1)) + \sum_{j=0}^{l} P_{l+1,j}(d(l+1))\mathcal{R}_j(d), \\
&\qquad l = 0, 1, \ldots, M-2; \\
\mathcal{Q}_{0,j}(d) &= P_{0,j}(d(0)); \\
\mathcal{Q}_{l+1,j}(d) &= \begin{cases} \sum_{i=0}^{l} [P_{l+1,i}(d(l+1))\mathcal{Q}_{i,j}(d)] & \text{for } j < l+1; \\ \sum_{i=0}^{l} [P_{l+1,i}(d(l+1))\mathcal{Q}_{i,j}(d)] + P_{l+1,j}(d(l+1)) & \text{for } j \geq l+1. \end{cases} \\
&\qquad l = 0, 1, \ldots, M-2.
\end{aligned}$$

In this case, the iterations (7) are similar to the Gauss-Seidel version of the value iteration algorithm (see [21, §6.3.3]). Assumption 1 is satisfied: the mapping $\hat{d} \in \mathbf{D}$ can be built consecutively for $s = 0, 1, \ldots, M-1$. This version of the general model appears in Section 7.

(iii) If the model $\langle \mathbf{D}, \mathcal{R}, \mathcal{Q} \rangle$ is fixed, one can construct the two-step model $\langle \bar{\mathbf{D}}, \bar{\mathcal{R}}, \bar{\mathcal{Q}} \rangle$ with $\bar{\mathbf{D}} := \mathbf{D} \times \mathbf{D}$, $\bar{\mathcal{R}}(\bar{d}) := \mathcal{R}(d_1) + \mathcal{Q}(d_1)\mathcal{R}(d_2)$ and $\bar{\mathcal{Q}}(\bar{d}) := \mathcal{Q}(d_1)\mathcal{Q}(d_2)$. If Assumption 1 is valid for the initial model, then it also holds for the two-step model. Similarly, one can introduce the three-step model and so on.

Lemma 1 *In the version (ii) of the general model, the matrix $\mathcal{Q}(d)$ is stochastic for all $d \in \mathbf{D}$, provided the original matrix $P(a)$ is stochastic for all $a \in \mathbf{A}$.*

For the proof see Lemma 2 of [20].

Definition 1

(a) A decision $d^{\beta *} \in \mathbf{D}$ is β-discounted optimal for $\beta \in (0,1)$ if

$$V^{\beta *} = U^\beta \circ V^{\beta *} = \max_{d \in \mathbf{D}} \left\{ \mathcal{R}(d) + \beta \mathcal{Q}(d) V^{\beta *} \right\} = \mathcal{R}(d^{\beta *}) + \beta \mathcal{Q}(d^{\beta *}) V^{\beta *}. \qquad (8)$$

Here $V^{\beta *} \in \mathbb{R}^M$ is the unique solution to the (left) equation above.

(b) A decision $d^*_B \in \mathbf{D}$ is Blackwell optimal if it is β-discounted optimal for all $\beta \in [\beta_0, 1)$ for some $\beta_0 \in [0,1)$.

(c) For each $d \in \mathbf{D}$,

$$\mathcal{Q}^*(d) := \lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} \mathcal{Q}^k(d)$$

is the 'limiting' [21] or 'stationary' [12] stochastic matrix;

$$D(d) := [I - \mathcal{Q}(d) + \mathcal{Q}^*(d)]^{-1} - \mathcal{Q}^*(d)$$

is the deviation matrix [12, 21].

(d) A decision $d^* \in \mathbf{D}$ is average optimal if, for all $d \in \mathbf{D}$,

$$\mathcal{Q}^*(d)\mathcal{R}(d) \le \mathcal{Q}^*(d^*)\mathcal{R}(d^*).$$

$\mathbf{D}^*$ *is the set of all average optimal decisions.*

All the introduced objects are well defined according to Proposition 1.

Let the time horizon in the model $\langle \mathbf{D}, \mathcal{R}, \mathcal{Q}\rangle$ be infinite and consider the average reward problem described as follows. On each step $t = 1, 2, \ldots$, the decision $d \in \mathbf{D}$ leads to the transition matrix $\mathcal{Q}(d)$ and, if the current state is $S(t-1) = s \in \mathbf{S}$, then the reward equals $\mathcal{R}_s(d)$. The target is to maximise the following objective

$$\lim_{T\to\infty} \frac{1}{T}\mathbb{E}_s^d\left[\sum_{t=1}^T \mathcal{R}_{S(t-1)}(d)\right] = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^T \mathcal{Q}^{t-1}_{s,\cdot}(d)\mathcal{R}(d) = [\mathcal{Q}^*(d)\mathcal{R}(d)]_s \qquad (9)$$
$$\to \max_{d\in\mathbf{D}}.$$

Here $S(t)$ is the (random) state at the time moment $t = 0, 1, 2, \ldots$ of the time-homogeneous Markov chain with the transition matrix $\mathcal{Q}(d)$ and the initial state $s = S(0) \in \mathbf{S}$; $\mathbb{E}_s^d$ is the corresponding mathematical expectation. The average (uniformly) optimal decision is that which provides the maximum in (9) simultaneously for all initial states $s \in \mathbf{S}$. In case of the standard model, we have the conventional MDP [12, 21], where only the stationary deterministic control strategies (policies) are considered; more general strategies (randomized, past-dependent) do not improve the gain. As will be shown (see Proposition 1(e)), the average optimal decision exists also in the general case.

In what follows, we will also deal with the discounted problem with $\beta \in (0, 1)$:

$$V^{\beta d}(s) := \lim_{T\to\infty}\mathbb{E}_s^d\left[\sum_{t=1}^T \beta^{t-1}\mathcal{R}_{S(t-1)}(d)\right] \to \max_{d\in\mathbf{D}} =: V^{\beta *}(s). \qquad (10)$$

Note that randomized, non-stationary and past-dependent decisions do not improve the objective: for the standard model see [12, 21], for the general case the proof is similar, based on the dynamic programming approach. The function $V^{\beta *}$ coincides with that introduced in Definition 1(a): see Proposition 1(b). All the reasoning from Section 2 remains the same, leading to the following version of the discounted Turnpike Theorem.

> For a fixed initial function $W^0 = W^{term}$, there exists the turnpike integer K^* such that, for all $k \ge K^*$, the inclusion
>
> $$\arg\max_{d\in\mathbf{D}}\{\mathcal{R}(d)+\beta\mathcal{Q}(d)W^k\} \subset \mathbf{D}^\beta_* := \{d \in \mathbf{D} :\ \mathcal{R}(d)+\beta\mathcal{Q}(d)V^{\beta *} = V^{\beta *}\}^{\cdot} \qquad (11)$$
>
> holds true.

Similarly to Section 2, for the finite horizon discounted problem, if the time horizon T is bigger than K^*, then on the first steps $t = 1, 2, \ldots, T - K^*$ one has to choose the decisions from $\mathbf{D}^\beta_*$.

Let the time horizon $T \in \{0, 1, 2, \ldots\}$ be finite and consider the undiscounted total reward problem with the terminal reward W^{term}:

$$\mathbb{E}_s^\pi\left[\sum_{t=1}^T \mathcal{R}_{S(t-1)}(d(t)) + W^{term}(S(T))\right] \to \max_{d(1),d(2),\ldots,d(T)\in\mathbf{D}} =: W^T(s).$$

Different decisions $d(1), d(2), \ldots, d(T)$ are allowed on the time steps $t = 1, 2, \ldots, T$, and π is the associated control strategy leading to the strategical measure $\mathbb{P}_s^\pi$ and to the corresponding mathematical expectation $\mathbb{E}_s^\pi$. Now $S(\cdot)$ is the (non-homogeneous) Markov chain. In case of the standard model, we have the conventional MDP, and it is well known that randomized and past-dependent decisions do not improve the objective [21, Prop.4.4.3].

Quite formally, one can choose different decisions $d(t)$ after observing the state $S(t-1)$. This concerns finite or infinite horizon cases, discounted or undiscounted reward. But according to the dynamic programming ideas and keeping in mind Assumption 1, one can restrict to the case when $d(t)$ is the same for all $S(t-1) \in \mathbf{S}$.

The Bellman function $W^T(s)$ satisfies equations (7) with $\beta = 1$, i.e.,

$$W^0 = W^{term} \quad \text{and} \quad W^{k+1} = U \circ W^k, \quad k = 0, 1, 2, \ldots, T-1.$$

The optimal decisions $d(t)$ on the steps $t = 1, 2, \ldots, T$ are those which provide the maximum in (7) at $k = T-1, T-2, \ldots, 0$.

In what follows, we use the 'span-seminorm'

$$sp(W) := \max_{s \in \mathbf{S}} W(s) - \min_{s \in \mathbf{S}} W(s)$$

in the space of (bounded) functions W on $\mathbf{S}$. The main result of the current section, Theorem 1, states that the condition $\lim_{k\to\infty} sp(W^{k+1} - W^k) = 0$ is sufficient for the turnpike property of the model. Note that in general the sequence $sp(W^{k+1} - W^k)$, $k = 0, 1, 2, \ldots$ does not approach zero [19, §4.2.18], [21, Ex.8.5.1]; see also examples below and in Subsection 7.4.

Theorem 1 (Turnpike) *Suppose for $W^0 = W^{term}$ the iterations (7) with $\beta = 1$ satisfy the requirement: $\lim_{k\to\infty} sp(W^{k+1} - W^k) = 0$. Then there exists K such that, in the described undiscounted model with the time horizon $T > K$, on the first steps $t = 1, 2, \ldots, T-K$ the optimal decisions are necessarily average optimal ones (in the sense of Definition 1(d)).*

Sufficient conditions for the requirement $\lim_{k\to\infty} sp(W^{k+1} - W^k) = 0$ are provided in Section 5. If this requirement is not satisfied, the turnpike property may be not valid: see Example below and Subsection 7.4.

Like in the discounted model, the Turnpike Theorem is useful if $\mathbf{D}^*$, the set of average optimal decisions, is a singleton. Sufficient conditions for this are given in the following statement.

Theorem 2 *Suppose for each $d \in \mathbf{D}$ the matrix $\mathcal{Q}(d)$ is irreducible, the Blackwell optimal strategy d_B is unique, and d^* is the unique decision providing the maximum in the optimality equation*

$$Ge + Y = \max_{d \in \mathbf{D}} \{\mathcal{R}(d) + \mathcal{Q}(d)Y\}. \tag{12}$$

(Under the imposed conditions, the vector Y is unique up to the additive constant.) Then $d^* = d_B$ is the unique average optimal decision.

Example. Here we show that the statement of Theorem 1 may be not valid if $sp(W^{k+1} - W^k)$ does not approach zero as $k \to \infty$.

Let $\mathbf{S} = \{0, 1, 2\}$; $\mathbf{A} = \{a_1, a_2, a_3\}$;

$$P_{0,1}(a_1) = P_{1,0}(a_1) = 1; \quad P_{0,1}(a_2) = P_{1,0}(a_2) = 1 - \varepsilon; \; P_{0,0}(a_2) = P_{1,1}(a_2) = \varepsilon;$$

$$P_{0,2}(a_3) = P_{1,2}(a_3) = P_{2,2}(a_1) = P_{2,2}(a_2) = P_{2,2}(a_3) = 1;$$

other transition probabilities equal zero. The one-step rewards $r_s(a)$ are given in the table:

a:	a_1	a_2	a_3
s:			
0	0	-h	1
1	0	-h	0
2	-2	-2	-2

We assume that

$$\varepsilon \in (0,1); \quad 0 < h < \varepsilon$$

and consider the standard model described in Item (i) above. Let $W^0 = W^{term} := \mathbf{0}$ and put $x_1 := 0$; $x_k := (1-\varepsilon)x_{k-1} + \varepsilon - h$ for $k \geq 2$.

Lemma 2

(a) $x_{k-1} < x_k < 1 - \frac{h}{\varepsilon}$ for all $k \geq 2$.

(b) If $k \geq 1$ is even, then $W^{k+1} = (1, x_{k+1}, -2(k+1))^T$ and the maximum in (7) is provided by the decision $d_o:\ d_o(0) = a_1,\ d_o(1) = a_2,\ d_o(2) = a_1$;
if $k \geq 1$ is odd, then $W^{k+1} = (x_{k+1}, 1, -2(k+1))^T$ and, for $k \geq 3$, the maximum in (7) is provided by the decision $d_e:\ d_e(0) = a_2,\ d_e(1) = a_1,\ d_e(2) = a_1$. The index o(e) corresponds to the odd (even) value of $k+1$.

In fact, the values $d_o(2)$ and $d_e(2)$ may be arbitrary because the state 2 is ultimately absorbing with the same one-step reward -2 for all actions.

Obviously, neither d_o, nor d_e is average optimal, but, if the time horizon $T \geq 3$ is odd (even), then the first optimal decision is d_o (d_e). One can show that the switching control strategies $(d_e, d_o, d_e, \ldots)$ and $(d_o, d_e, d_o, \ldots)$ are both average optimal. If the time horizon T is finite without the terminal reward, it is desirable to try to reach the state $S(T-1) = 0$ and apply the action a_3 in order to gain $r_0(a_3) = 1$ at the last step. If T is even, then one can consider the two-step model $\langle \bar{\mathbf{D}}, \bar{\mathcal{R}}, \bar{\mathcal{Q}} \rangle$ as in Item (iii) below Assumption 1. Now the decision $\bar{d}^* := (d_e, d_o)$ is average optimal, and it should be applied always up to the last pair of time steps, that is, the turnpike property is valid.

4 Proposition 1 and Proof of Theorems 1 and 2

Proposition 1

(a) A β-discounted optimal decision exists for each $\beta \in (0,1)$.

(b) Suppose $\beta \in (0,1)$ is fixed. For each $d \in \mathbf{D}$, the unique solution to the equation

$$V^{\beta d} = \mathcal{R}(d) + \beta \mathcal{Q}(d) V^{\beta d} \tag{13}$$

satisfies the inequality $V^{\beta d} \leq V^{\beta *}$.

(c) For each $d \in \mathbf{D}$ the limiting matrix $\mathcal{Q}^*(d)$ exists, is stochastic, and exhibits the following properties:

$$\mathcal{Q}^*(d) = \mathcal{Q}(d)\mathcal{Q}^*(d) = \mathcal{Q}^*(d)\mathcal{Q}(d) = \mathcal{Q}^*(d)\mathcal{Q}^*(d).$$

(d) The deviation matrix $D(d)$ is well defined and satisfies equation

$$D(d) = [I - \mathcal{Q}(d) + \mathcal{Q}^*(d)]^{-1}](I - \mathcal{Q}^*(d)).$$

(e) There exists a Blackwell optimal decision, and every Blackwell optimal decision is also average optimal. Hence an average optimal decision also exists and $\mathbf{D}^* \neq \emptyset$.

Lemma 3 *Suppose a decision $\hat{d} \in \mathbf{D}$ provides maximum to*

$$\mathcal{R}(d) + \mathcal{Q}(d)W,$$

where $W \in \mathbb{R}^M$. Then

$$\begin{aligned} &\min_{s \in \mathbf{S}} \{U \circ W(s) - W(s)\} e \leq \mathcal{Q}^*(\hat{d})\mathcal{R}(\hat{d}) \leq \mathcal{Q}^*(d^*)\mathcal{R}(d^*) \\ \leq\ &\max_{s \in \mathbf{S}} \{U \circ W(s) - W(s)\} e, \end{aligned}$$

where d^* is an average optimal decision.

Proof of Theorem 1. Denote $\mathbf{D}^*$ the set of all average optimal decisions in the model $\langle \mathbf{D}, \mathcal{R}, \mathcal{Q} \rangle$. This set is not empty due to Proposition 1(e). Let $g^* := \mathcal{Q}^*(d^*)\mathcal{R}(d^*)$ for $d^* \in \mathbf{D}^*$. Clearly, the vector g^* does not depend on the choice of $d^* \in \mathbf{D}^*$. Let

$$\varepsilon := \min_{d \in \mathbf{D} \setminus \mathbf{D}^*} \|g^* - \mathcal{Q}^*(d)\mathcal{R}(d)\| > 0,$$

where, as usual, $\|V\| := \max_{s \in \mathbf{S}} |V(s)|$ is the uniform norm in $\mathbb{R}^M$. Take K such that

$$sp(W^{k+1} - W^k) = sp(U \circ W^k - W^k) < \varepsilon$$

for all $k \geq K$.

Now, for each $k \geq K$, if $\hat{d} \in \mathbf{D}$ provides the maximum to $\mathcal{R}(d) + \mathcal{Q}(d)W^k$ then, by Lemma 3,

$$\begin{aligned} &g^* - \mathcal{Q}^*(\hat{d})\mathcal{R}(\hat{d}) = \mathcal{Q}^*(d^*)\mathcal{R}(d^*) - \mathcal{Q}^*(\hat{d})\mathcal{R}(\hat{d}) \\ \leq\ &sp(U \circ W^k - W^k)e < \varepsilon e \Longrightarrow \|g^* - \mathcal{Q}^*(\hat{d})\mathcal{R}(\hat{d})\| < \varepsilon; \end{aligned}$$

hence $\hat{d} \in \mathbf{D}^*$.

Therefore, for each $T > K$, on the first steps $t = 1, 2, \dots, T-K$, the optimal decisions, which provide the maximum to $\mathcal{R}(d) + \mathcal{Q}(d)W^k$ at $k = T-1, T-2, \dots, K$, belong to $\mathbf{D}^*$. □

Proof of Theorem 2. If the matrix $\mathcal{Q}(d)$ is irreducible, then the stochastic matrix $\mathcal{Q}^*(d)$ has identical strictly positive rows: see Thm.5.1.1 and Thm.5.1.4 of [13]. Therefore, $\mathcal{Q}^*(d)\mathcal{R}(d) = G(d)e$ for some $G(d) \in \mathbb{R}$ and $d^* \in \mathbf{D}^*$ if and only if $G(d^*) = \max_{d \in \mathbf{D}} G(d) =: G^*$. Recall also that a Blackwell optimal decision exists according to Proposition 1(e).

The optimality equation (12) is solvable, and a pair $(G, Y) \in \mathbb{R} \times \mathbb{R}^M$ satisfies it if and only if

$$G = G^* \quad \text{and } Y = ce + D(d_B)\mathcal{R}(d_B),$$

where $c \in \mathbb{R}$ is an arbitrary constant, d_B is some Blackwell optimal decision and $D(d_B)$ is the deviation matrix. In the case of the standard model, this is precisely Theorem 6.1 of [12]; all the steps of its proof remain correct for the general case.

Let us show that a decision d^* is average optimal if and only if it provides the maximum in (12). Let (G^*, Y) be a solution to the equation (12). Then

$$\mathcal{G}(d) := \mathcal{R}(d) - G^*e + \mathcal{Q}(d)Y - Y \leq \mathbf{0} \text{ for all } d \in \mathbf{D}, \quad \text{and } \max_{d \in \mathbf{D}} \mathcal{G}(d) = \mathbf{0}.$$

Since all the rows of each matrix $\mathcal{Q}^*(d)$ are strictly positive,

$$\mathcal{Q}^*(d)\mathcal{G}(d) \leq \mathbf{0} \text{ for all } d \in \mathbf{D} \quad \text{and } \max_{d \in \mathbf{D}} \mathcal{Q}^*(d)\mathcal{G}(d) = \mathbf{0}.$$

Moreover, for a decision $\hat{d}$ providing the last maximum, we have

$$\mathcal{Q}^*(\hat{d})\mathcal{G}(\hat{d}) = \max_{d\in\mathbf{D}} \mathcal{Q}^*(d)\mathcal{G}(d) = \mathbf{0} \Longleftrightarrow \mathcal{G}(\hat{d}) = \max_{d\in\mathbf{D}} \mathcal{G}(d) = \mathbf{0} \tag{14}$$

because $\mathcal{Q}^*(d) > 0$ for all $d \in \mathbf{D}$. But $\mathcal{Q}^*(d)\mathcal{G}(d) = \mathcal{Q}^*(d)\mathcal{R}(d) - G^*e$ for all $d \in \mathbf{D}$ (see Proposition 1(c)), so that a decision d^* is average optimal if and only if

$$\mathbf{0} = \mathcal{Q}^*(d^*)\mathcal{G}(d^*) = \max_{d\in\mathbf{D}} \mathcal{Q}^*(d)\mathcal{G}(d) \Longleftrightarrow \mathcal{G}(d^*) = \max_{d\in\mathbf{D}} \mathcal{G}(d) \text{ by (14)},$$

i.e., if and only if d^* provides the maximum in (12).

Now the statement of Theorem 2 follows. Recall that the decision d_B is average optimal by Proposition 1(e), so that $d^* = d_B$. □

5 Sufficient Conditions for Equation $\lim_{k\to\infty} sp(W^{k+1} - W^k) = 0$ in Case $\beta = 1$

For fixed $J \geq 1$ take arbitrary sequences $d^1 = \{d_1^1, d_2^1, \ldots, d_J^1\}$ and $d^2 = \{d_1^2, d_2^2, \ldots, d_J^2\}$ of decisions from $\mathbf{D}$ and matrices

$$\begin{aligned} \mathcal{Q}^{(J,1)} &:= \mathcal{Q}(d_1^1)\mathcal{Q}(d_2^1)\ldots\mathcal{Q}(d_J^1); \\ \mathcal{Q}^{(J,2)} &:= \mathcal{Q}(d_1^2)\mathcal{Q}(d_2^2)\ldots\mathcal{Q}(d_J^2). \end{aligned}$$

Denote

$$\eta(d^1, d^2) := \min_{(s,u)\in\mathbf{S}^2} \sum_{l\in\mathbf{S}} \min\left\{\mathcal{Q}_{s,l}^{(J,1)}, \mathcal{Q}_{u,l}^{(J,2)}\right\}$$

and introduce

$$\gamma^J := 1 - \min_{(d^1,d^2)} \eta(d^1, d^2) \in [0, 1]. \tag{15}$$

Lemma 4 *Suppose $\gamma^J < 1$ for some $J \geq 1$. Then the following statements hold.*

(a) The operator $U := U^1$ in (7) is a J-step contraction with respect to the span-seminorm:

$$sp(U^J \circ V_1 - U^J \circ V_2) \leq \gamma^J sp(V_1 - V_2)$$

for all $V_1, V_2 \in \mathbb{R}^M$.

(b) For every $W^0 \in \mathbb{R}^M$ $\lim_{k\to\infty} sp(W^{k+1} - W^k) = 0$.

Remark 2 *For the standard model, the statement of Theorem 1 under the condition $\gamma^1 < 1$ was formulated in [14] without proof: see Condition A and the main theorem therein.*

Sufficient conditions for the inequality $\gamma^J < 1$ at some $J \geq 1$ are given in the following lemma.

Lemma 5 *Suppose there exist $J \geq 1$ and a state $l \in \mathbf{S}$ such that, for any sequence $\{d_1, d_2, \ldots, d_J\}$ of decisions from $\mathbf{D}$, for the matrix*

$$\mathcal{Q}^{(J)} := \mathcal{Q}(d_1)\mathcal{Q}(d_2)\ldots\mathcal{Q}(d_J),$$

the strict inequality $\mathcal{Q}_{s,l}^{(J)} > 0$ is valid for all $s \in \mathbf{S}$.

Then $\gamma^J < 1$, and hence $\lim_{k\to\infty} sp(W^{k+1} - W^k) = 0$ for every $W^0 \in \mathbb{R}^M$.

The proof is identical to the proof of Theorem 8.5.3(b) of [21].

Corollary 1 *If, for each $d \in \mathbf{D}$, the matrix $\mathcal{Q}(d)$ is irreducible and aperiodic and $\mathcal{Q}_{s,s}(d) > 0$ for all $s \in \mathbf{S}$, then $\gamma^J < 1$ for some $J \geq 1$, and hence $\lim_{k\to\infty} sp(W^{k+1} - W^k) = 0$ for every $W^0 \in \mathbb{R}^M$.*

Lemma 6 *Suppose for every average optimal decision $d^* \in \mathbf{D}^*$ the matrix $\mathcal{Q}(d^*)$ is aperiodic. Assume also that $\mathcal{Q}^*(d^*)\mathcal{R}(d^*) = G^*e$: the maximal value of the objective (9) does no depend on the initial state s. Then, for every $W^0 \in \mathbb{R}^M$, $\lim_{k\to\infty} sp(W^{k+1} - W^k) = 0$.*

Remark 3 *Suppose the model is standard.*

- *The turnpike theorem presented in [17] follows from Lemma 6.*
- *Sufficient conditions for the turnpike property as in [8] are similar to the conditions in Lemma 6 and Theorem 2. (See [8, Thm.4.4].)*

6 Discounted Model

Suppose the function $W^0 = W^{term} : \mathbf{S} \to \mathbb{R}$ is fixed and the statement of Theorem 1 is valid. Let us fix an *arbitrary* $K_1 > K + 1$ and let the time horizon satisfy $K < T < K_1$. On the first steps $t = 1, 2, \dots, T - K$ the optimal undiscounted decisions, i.e., those which provide the maximum in (7) with $\beta = 1$ at $k = T - 1, T - 2, \dots, K$, necessarily belong to $\mathbf{D}^*$, that is, are average optimal. Denote $W^{\beta k}$ the functions from (7) with $\beta \in (0, 1)$ and with the same initial condition $W^{\beta 0} = W^{term}$. Obviously, for each $k = 1, 2, \dots, K_1 - 2$ the vector $W^{\beta k} \in \mathbb{R}^M$ is continuous with respect to β. (Recall that $\mathbf{S} = \{0, 1, \dots, M - 1\}$.) Since the spaces $\mathbf{S}$ and $\mathbf{A}$ are finite, for all β close enough to 1 the optimal decisions on the first steps $t = 1, 2, \dots, T - K$ in the β-discounted model again necessarily belong to $\mathbf{D}^*$. (Recall that $T < K_1$.) One can rephrase this observation in the following way.

> For an arbitrarily fixed $K_1 > K + 1$ there exists $\beta_0 \in (0, 1)$ such that for each fixed $\beta \in [\beta_0, 1]$, in the β-discounted model $\langle \mathbf{D}, \mathcal{R}, \mathcal{Q} \rangle$ with the fixed terminal reward W^{term} and an arbitrarily fixed time horizon $T \geq K_1$, if the remaining number of steps is between K and K_1, then the optimal decisions necessarily belong to $\mathbf{D}^*$.

Roughly speaking, for such big values of β one can ignore the discount factor on the steps $t = T - K_1 + 2, T - K_1 + 3, \dots, T - K$. See Fig.1.

On the other hand, for that value of $\beta \in [\beta_0, 1)$, as explained in Section 3, there exists the turnpike integer K^*: see (11). If the time horizon $T > K^*$ then on the first steps $t = 1, 2, \dots, T - K^*$ the optimal decisions necessarily belong to $\mathbf{D}^\beta_*$, the set of optimal decisions in the (infinite horizon) discounted model $\langle \mathbf{D}, \mathcal{R}, \mathcal{Q} \rangle$. See Fig.1.

To summarize, if the statement of Theorem 1 is valid, then, for $\beta \approx 1$, there can appear *two* turnpikes as on Fig.1. When $\beta \to 1-$, one can take larger and larger K_1, i.e., $K_1 \to \infty$, and in the limiting case, when $\beta = 1$, we have just the turnpike 1 as in Theorem 1. Note also that, for β close enough to 1, $\mathbf{D}^\beta_*$ contains only Blackwell optimal decisions, hence $\mathbf{D}^\beta_* \subset \mathbf{D}^*$.

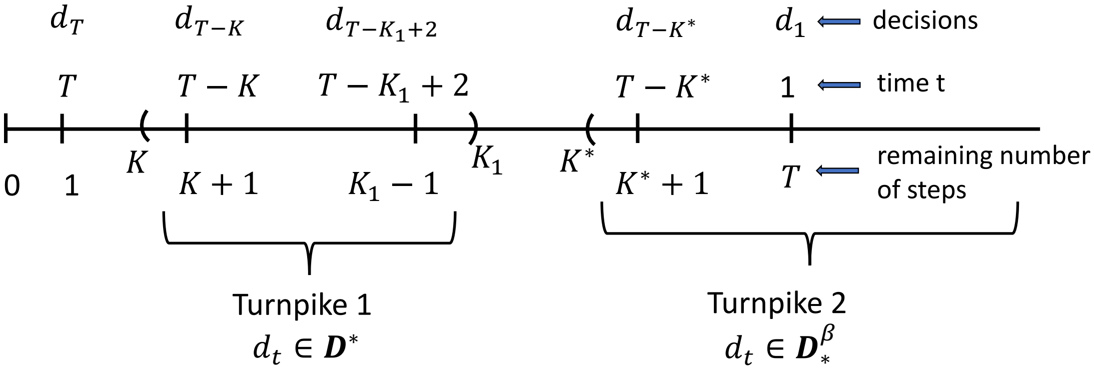

Figure 1: Turnpikes for $\beta \approx 1$.

7 Controlled Random Walk

7.1 Description of the Model and Associated Conjecture

The problem under study is a specific Markov Decision Process (MDP) defined as follows.

Suppose

$$\mathbf{A} = \{a_1, a_2, \dots, a_N\}$$

is the finite action space and, for each $a \in \mathbf{A}$, let $Z(a)$ be the random variable taking values $m = 1, 2, \dots, M$ with probabilities $p_m(a)$. The state space of the MDP is

$$\mathbf{X} := \{-M, -M+1, -M+2, \dots\}.$$

All the states $-M, -M+1, \dots, -1$ are absorbing, with zero rewards.

If action $a \in \mathbf{A}$ is chosen in the state $i \geq 0$, then the new state is just $j = i - Z(a)$; another choice of action $a \in \mathbf{A}$ leads to the independent random variable $Z(a)$. Thus, the transition probability is given by

$$\tilde{P}_{i,j}(a) = \begin{cases} p_m(a), & \text{if } j = i - m, \quad m = 1, 2, \dots, M; \\ 0 & \text{otherwise.} \end{cases}$$

See Fig.2.

Finally, the associated (expected) reward in states $i \geq 0$ equals R^a and is i-independent. For example, if $R_{Z(a)}(a)$ is the reward associated with the action $a \in \mathbf{A}$ and the value $Z(a)$, then

$$R^a = \sum_{m=1}^{M} R_m(a) p_m(a).$$

To summarise, the one-step rewards in the MDP equal

$$r_i(a) := \begin{cases} 0, & \text{if } i < 0; \\ R^a, & \text{if } i \geq 0. \end{cases}$$

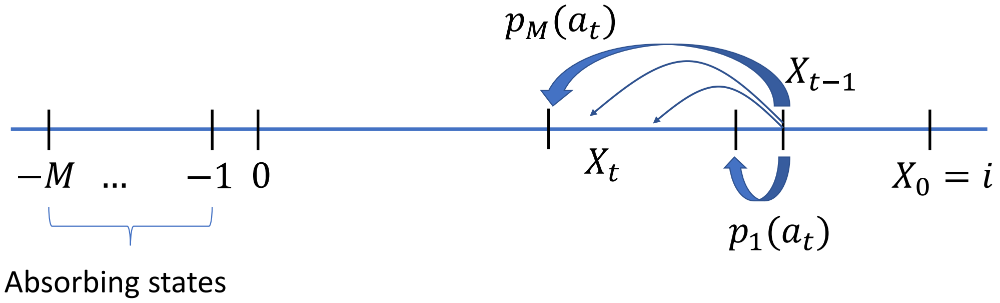

Figure 2: Random walk.

The initial state $i \in \mathbf{X}$ is fixed, and we consider MDP $\langle \mathbf{X}, \mathbf{A}, \tilde{P}, r\rangle$ with the total expected reward, with (random) states and actions

$$X_0 = i,\ A_1,\ X_1,\ A_2, \dots\ .$$

The definition of a strategy π (past-dependent, randomized) is conventional [12, 18, 21]; E_i^π is the corresponding mathematical expectation;

$$V(i) := \sup_\pi \mathbb{E}_i^\pi \left[\sum_{t=1}^\infty r_{X_{t-1}}(A_t)\right] \tag{16}$$

is the Bellman function for this MDP; $i \in \mathbf{X}$. Since the reward r is bounded and the process X_t is ultimately absorbed after (maximum) $i+1$ time steps at $\{-M, -M+1, \dots, -1\}$, the function V is finite-valued. It is well known (see, e.g., [12, Ch.4] or [13, §9.5]) that the function V is the unique solution to the optimality (Bellman) equation

$$\begin{aligned} V(i) &= \max_{a\in\mathbf{A}}\left\{R^a + \sum_{m=1}^M V(i-m)p_m(a)\right\} \quad \text{for } i \ge 0; \\ V(i) &= 0 \quad \text{for } i = -M, -M+1, \dots, -1, \end{aligned} \tag{17}$$

which can be solved successively for $i = 0, 1, \dots$.

The described random walk is a version of the adaptive stochastic knapsack problem, see [1, 2]. Indeed, i is the capacity of the knapsack, $\mathbf{A}$ is the set of types of items. Each type has infinite supply of items. Type $a \in \mathbf{A}$ has value R^a and the random size $Z(a)$. In our setting, the first item which makes the knapsack broken (i.e., the total size of items in the knapsack exceeds the capacity) is still counted. One can easily modify the initial values $V(i)$, $i = 0, 1, \dots, M-1$ in such a way that it is not taken into account:

$$V(i) = \max_{a\in\mathbf{A}}\left\{\sum_{m=1}^i p_m(a)[R^a + V(i-m)]\right\} \text{ for } i \le M-1.$$

I am thankful to Prof.N.Bäuerle for pointing this interpretation of the random walk.

Let us introduce the following notations: $L^a := \mathbb{E}[Z(a)] = \sum_{m=1}^{M} mp_m(a) \geq 1$; $c_* := \max_{a\in\mathbf{A}} \frac{R^a}{L^a}$, and

$$\mathbf{A}_* := \{a \in \mathbf{A} : \frac{R^a}{L^a} = c_*\}. \tag{18}$$

For a fixed $a \in \mathbf{A}$, for a large initial state i, the total expected number of time steps up to the absorption equals $\approx \frac{i}{L^a}$. Thus, the total expected reward is $\approx i\frac{R^a}{L^a}$. To put it slightly different, the expected reward, coming from passing through one point on the lattice, equals $\approx \frac{R^a}{L^a}$. In the current section, we show that, under appropriate conditions, the following **conjecture** is valid

$$\begin{array}{c}\text{There is such } I < \infty \text{ that, for all } i \geq I,\\ \text{the maximum in (17) is only provided by } a \in \mathbf{A}_* :\end{array} \tag{19}$$

see Theorem 3. This conjecture was formulated by Prof.I.Sonin in a private conversation. Example in Subsection 7.4 shows that this conjecture may be not valid in some situations.

Remark 4 *One can notice that equation (17) is a special case of the equation (1) of [16]. In that article, no conjectures like (19) were investigated. The main results concerned the behaviour of the function V: for example, under mild conditions, for $\Delta V(i) := V(i) - V(i-1)$, the sequence* $\max_{1\leq m\leq M} \Delta V(i-m+1)$ *converges as $i \to \infty$.*

7.2 Proof of the Conjecture

We will reformulate the conjecture (19) as the Turnpike Theorem provided in Section 3. The first step is described in the following lemma.

Lemma 7 *The function*

$$\tilde{W}(i) := V(i) - c_* i, \quad i \in \mathbf{X} \tag{20}$$

is the (unique) uniformly bounded function satisfying equation

$$\begin{aligned} \tilde{W}(i) &= -c_* i \quad \textit{for } i = -M, -M+1, \dots, -1; \\ \tilde{W}(i) &= \max_{a\in\mathbf{A}} \left\{ L^a \left(\frac{R^a}{L^a} - c_* \right) + \sum_{m=1}^{M} \tilde{W}(i-m)p_m(a) \right\} \quad \textit{for } i \geq 0, \end{aligned} \tag{21}$$

which can be solved successively for $i = 0, 1, \dots$. Hence $V(i) = c_* i + O(1)$ when $i \to \infty$.

Moreover, for each $i \in \mathbf{X}$, the maxima in (17) and in (21) are provided by the same values of $a \in \mathbf{A}$.

For the proof see Lemma 1 of [20].

Every value of $i \in \mathbf{X}$ can be uniquely represented as

$$i = (k-1)M + s, \quad \text{where } s \in \mathbf{S} := \{0, 1, \dots, M-1\}, \quad k = 0, 1, 2, \dots.$$

For each $i \in \mathbf{X}$ with the corresponding values of k and s, we denote $\tilde{W}(i)$, introduced in (20), as $W^k(s)$. Now equation (21) takes the following form:

$$W^0(s) = -c_*(-M + s) \quad \text{for } s \in \mathbf{S};$$

$$\begin{aligned}
W^{k+1}(0) &= \max_{a\in\mathbf{A}}\left\{L^a\left(\frac{R^a}{L^a}-c_*\right)+\sum_{j=0}^{M-1}W^k(j)p_{M-j}(a)\right\},\\
W^{k+1}(1) &= \max_{a\in\mathbf{A}}\left\{L^a\left(\frac{R^a}{L^a}-c_*\right)+W^{k+1}(0)p_1(a)+\sum_{j=1}^{M-1}W^k(j)p_{M-j+1}(a)\right\},\\
\dots\\
W^{k+1}(M-1) &= \max_{a\in\mathbf{A}}\left\{L^a\left(\frac{R^a}{L^a}-c_*\right)+W^{k+1}(M-2)p_1(a)+W^{k+1}(M-3)p_2(a)\right.\\
&\quad\left.+\dots+W^{k+1}(0)p_{M-1}(a)+W^k(M-1)p_M(a)\right\},\quad k=0,1,\dots.
\end{aligned}$$

After we introduce the stochastic matrix

$$P(a):=\left(\begin{array}{llll}
P_{0,0}(a)=p_M(a) & P_{0,1}(a)=p_{M-1}(a) & \dots & P_{0,M-1}(a)=p_1(a)\\
P_{1,0}(a)=p_1(a) & P_{1,1}(a)=p_M(a) & \dots & P_{1,M-1}(a)=p_2(a)\\
\dots & & \dots & \\
P_{M-1,0}(a)=p_{M-1}(a) & P_{M-1,1}(a)=p_{M-2}(a) & \dots & P_{M-1,M-1}(a)=p_M(a)
\end{array}\right), \tag{22}$$

the obtained equations for $W^k(s)$ can be rewritten as

$$\begin{aligned}
W^0(s) &= -c_*(-M+s)\quad \text{for } s\in\mathbf{S}; &(23)\\
W^{k+1}(s) &= \max_{a\in\mathbf{A}}\left\{L^a\left(\frac{R^a}{L^a}-c_*\right)+\sum_{j=0}^{s-1}W^{k+1}(j)P_{s,j}(a)+\sum_{j=s}^{M-1}W^k(j)P_{s,j}(a)\right\}\\
&\quad \text{for } s\in\mathbf{S},\ k\ge 0.
\end{aligned}$$

Let us put $r_s(a):=L^a\left(\frac{R^a}{L^a}-c_*\right)\le 0$ and consecutively for $s=0,1,2,\dots,M-1$ express $W^{k+1}(s)$ in terms of W^k. As the result, we finish with the equations

$$\begin{aligned}
W^0(s) &= -c_*(-M+s),\quad s\in\mathbf{S};\\
W^{k+1} &= \max_{d\in\mathbf{D}}\{\mathcal{R}(d)+\mathcal{Q}(d)W^k\},\quad k=0,1,\dots,
\end{aligned}\tag{24}$$

where $\mathbf{D}$ is the set of all mappings $d:\ \mathbf{S}\to\mathbf{A}$, and $\mathcal{R}(d)\le\mathbf{0}$ and $\mathcal{Q}(d)$ are as in the version (ii) of the general model $\langle\mathbf{D},\mathcal{R},\mathcal{Q}\rangle$ described in Section 3.

Definition 2 *Decisions $d\in\mathbf{D}$ satisfying the property $d(s)\in\mathbf{A}_*$ for all $s\in\mathbf{S}$ will be called trivial. Equivalently, a decision $d\in\mathbf{D}$ is trivial if and only if $\mathcal{R}(d)=\mathbf{0}$.*

The conjecture (19) is now reformulated as follows:

There exists K such that, for all $k\ge K$, the maximum in (24) is only provided by the trivial decisions. (25)

Note that all the vectors $W^0,W^1,\dots$ are uniformly bounded by Lemma 7, and the maxima in (17),(21) and (23) are provided by the same values of $a\in\mathbf{A}$.

Since $\mathcal{R}(d)\le\mathbf{0}$ for all $d\in\mathbf{D}$ and $\mathcal{R}(d)=\mathbf{0}$ for each trivial decision d, it is obvious that $\mathbf{D}^*$, the set of average optimal decisions in the model $\langle\mathbf{D},\mathcal{R},\mathcal{Q}\rangle$, contains all trivial decisions: the optimal gain $\max_{d\in\mathbf{D}}\mathcal{Q}^*(d)\mathcal{R}(d)=\mathbf{0}$ is attained at any trivial decision.

Suppose the iterations (24) satisfy the requirement $\lim_{k\to\infty}sp(W^{k+1}-W^k)=0$. Then, according to the Turnpike Theorem 1, there exists K such that for each $T>K$ the maximum in (24) at $k=T-1$ is necessarily provided by an average optimal decision $d\in\mathbf{D}^*$ in the model $\langle\mathbf{D},\mathcal{R},\mathcal{Q}\rangle$. If $\mathbf{D}^*$ contains only the trivial decisions, then the conjecture (25) (and also (19)) is valid. This observation makes possible to formulate the sufficient conditions for the conjecture (19) in terms of the matrices $\mathcal{Q}(d)$ introduced above and based on the matrix (22).

Condition 1 *For each $d^* \in \mathbf{D}^*$ the matrix $\mathcal{Q}(d^*)$ corresponds to the Markov chain without transient states.*

Lemma 8 *If Condition 1 is satisfied then $\mathbf{D}^*$ contains only trivial decisions.*

Condition 2 *Either*

- *$\gamma^J < 1$ for some $J \geq 1$, where γ^J is as defined at the beginning of Section 5,*
- *or, for each $d^* \in \mathbf{D}^*$, the matrix $\mathcal{Q}(d^*)$ is aperiodic.*

Theorem 3 *If Conditions 1 and 2 are satisfied, then Conjecture (19) is valid.*

The proof directly follows from Lemmas 4, 6, Theorem 1 and Lemma 8. See also the explanations above.

In the following corollary, we provide the sufficient conditions for Conjecture (19) to be valid, in terms of the original random walk.

Corollary 2 *Suppose $p_M(a) > 0$ for all $a \in \mathbf{A}$ and, for any two states $i, j \in \mathbf{S}$, there exists a path $i_0 = i \to i_1 \to \ldots \to i_N = j$ in $\mathbf{S}$ such that, for any $a_0, a_1, \ldots, a_{N-1} \in \mathbf{A}$,*

$$P_{i_0,i_1}(a_0)P_{i_1,i_2}(a_1)\ldots P_{i_{N-1},i_N}(a_{N-1}) > 0,$$

where $P(a)$ is given by (22).

Then Conjecture (19) is valid.

The matrix $P(a)$ has a cyclic structure. Thus, the conditions of Corollary 2 are satisfied if there is $m < M$ having no common divisors with M such that $p_M(a) > 0$ and $p_m(a) > 0$ for all $a \in \mathbf{A}$.

In [20], the investigated random walk was studied using another method of attack and leading to the following statements.

Proposition 2 *Suppose there exists J such that, for every sequence $\hat{d}_1, \hat{d}_2, \ldots, \hat{d}_J$ of trivial decisions, the matrix $\mathcal{Q}(\hat{d}_1)\mathcal{Q}(\hat{d}_2)\ldots\mathcal{Q}(\hat{d}_J)$ contains no zeroes. Then Conjecture (19) is valid.*

For the proof see Theorem 1 of [20]. By the way, during the proof of that theorem, it was shown that there is a finite limit $\lim_{i\to\infty} \tilde{W}(i)$, so that $\lim_{i\to\infty}[V(i) - V(i-1)] = c_*$. (Cf. Remark 4.)

Corollary 3 *Suppose $\mathbf{A}_* = \{a_*\}$ is a singleton. (Consequently, there is a unique trivial decision $d(s) \equiv a_*$.) Let the matrix $P(a_*)$ be irreducible and aperiodic and assume that $p_M(a_*) > 0$. Then the conjecture (19) is valid.*

For the proof see Corollary 1 of [20].

7.3 Discounted Model

In this section, $\beta \in (0,1)$ is the discount factor, expression (16) is replaced with

$$V^\beta(i) := \sup_\pi \mathbb{E}_i^\pi \left[\sum_{t=1}^{\infty} \beta^{t-1} r_{X_{t-1}}(A_t)\right], \tag{26}$$

and the optimality equation looks like

$$\begin{aligned} V^\beta(i) &= \max_{a\in\mathbf{A}} \left\{ R^a + \beta \sum_{m=1}^{M} V^\beta(i-m)p_m(a) \right\} \quad \text{for } i \geq 0; \qquad (27)\\ V^\beta(i) &= 0 \quad \text{for } i = -M, -M+1, \ldots, -1. \end{aligned}$$

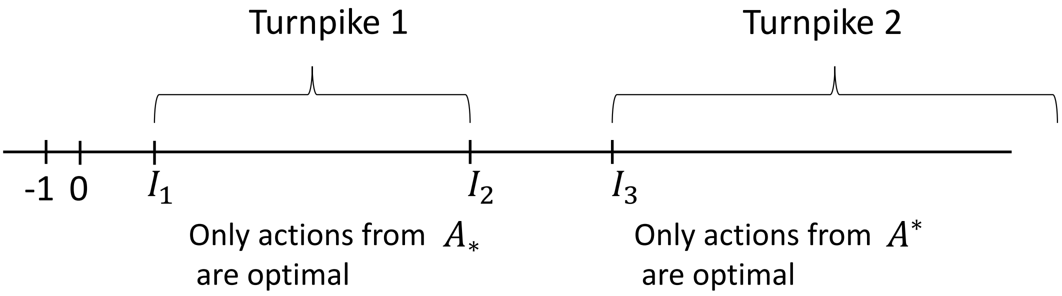
Turnpike 1
Turnpike 2
-1
0
I_1
I_2
I_3
Only actions from A_* are optimal
Only actions from A^* are optimal

Figure 3: Turnpikes for $\beta \approx 1$.

Like previously, it can be solved successively for $i = 0, 1, \ldots$. We put $R^* := \max_{a\in\mathbf{A}} R^a$ and

$$\mathbf{A}^* := \{a \in \mathbf{A} : \; R^a = R^*\}. \tag{28}$$

In [20] the following turnpike theorem was proved.

Theorem 4 *There is such $I < \infty$ that, for all $i \geq I$, the maximum in (27) is only provided by $a \in \mathbf{A}^*$.*

Remark 5 *Note that, if we transform this random walk to the model $\langle \mathbf{D}, \mathcal{R}, \mathcal{Q}\rangle$ as it was done in Subsection 7.2, we will not finish with the discounted problem (10) leading to iterations (7): the discount factor β appears in the calculation of the vectors $\mathcal{R}(d)$.*

It is interesting to look at what happens if the discount factor β is close to 1, assuming that Conjecture (19) is valid for $\beta = 1$. Denote the corresponding I as I_1, and fix an *arbitrary* $I_2 > I_1$. Obviously, for each $i \in \mathbf{X}$, $\lim_{\beta\to 1-} V^\beta(i) = V(i)$ with V as in (17). Therefore there exists $\beta_0 \in (0,1)$ such that, for all $i = I_1, I_1+1, \ldots, I_2$, for all $\beta \in [\beta_0, 1]$

$$\max_{a\in\mathbf{A}} \left\{ R^a + \sum_{m=1}^{M} V^\beta(i-m)p_m(a) \right\} > \max_{a\in\mathbf{A}\setminus\mathbf{A}_*} \left\{ R^a + \sum_{m=1}^{M} V^\beta(i-m)p_m(a) \right\}.$$

Thus, for a fixed $\beta \in [\beta_0, 1]$, for all $i = I_1, I_1+1, \ldots, I_2$, the maximum in (27) is only provided by $a \in \mathbf{A}_*$. Of course, by Theorem 4, there is a finite $I_3 > I_2$ such that, for all $i \geq I_3$, the maximum in (27) is only provided by $a \in \mathbf{A}^*$. Recall that $\mathbf{A}_*$ and $\mathbf{A}^*$ are given by (18) and (28) correspondingly. One can say that, for β close to 1, there are two turnpikes, where only actions from $\mathbf{A}_*$ and $\mathbf{A}^*$ are optimal in the model (26): see Fig.3. When β approaches 1, I_2 and I_3 go to infinity, and in the limiting case $\beta = 1$ we have just the conjecture (19).

The situation is similar to Section 6, but the turnpikes 2 are absolutely different: for any value of $\beta \in (0,1)$, here the actions in the second turnpike are usually not Blackwell optimal, and $\mathbf{D}_*^\beta$ contains only Blackwell optimal decisions if β is close enough to 1. (See also Remark 5.)

7.4 Example

In this subsection, we show that the conjecture (19) may not be valid if Conditions 1 and 2 or conditions of Proposition 2 and Corollary 3 are not satisfied. We use the notations from Subsection 7.1. In particular, the function V comes from the iterations (17).

Put

$$\mathbf{A} := \{a_1, a_2\},\ M := 3,\ \varepsilon \in (0,1),\ p_2(a_1) = 1,\ p_2(a_2) = 1-\varepsilon,\ p_3(a_2) = \varepsilon,$$

where $\varepsilon \in (0,1)$; other probabilities being zero. Finally, let $R^{a_1} := 2$ and $R^{a_2} := h \in (2, 2+\varepsilon)$. Now

$$L^{a_1} = 2,\ L^{a_2} = 2+\varepsilon,\ \frac{R^{a_1}}{L^{a_1}} = 1,\ \frac{R^{a_2}}{L^{a_2}} = \frac{h}{2+\varepsilon} < 1,\ c_* = 1,\ \mathbf{A}_* = \{a_1\}.$$

Since $h > 2$, obvious calculations lead to the following expressions:
$V(0) = V(1) = \max\{2, h\} = h$;
$V(2) = \max\{2 + V(0) = 2 + h;\ \ h + (1-\varepsilon)V(0) + \varepsilon V(-1) = h + (1-\varepsilon)h\} = 2 + h$
because

$$\frac{2}{1-\varepsilon} = 2[1 + \varepsilon + \varepsilon^2 + \ldots] > 2 + \varepsilon > h \Longrightarrow 2 > (1-\varepsilon)h.$$

$V(3) = \max\{2 + V(1) = 2 + h;\ \ h + (1-\varepsilon)V(1) + \varepsilon V(0) = 2h\} = 2h$. Further properties of the function V are given in the following lemma.

Lemma 9 *For all $j \geq 1$, the following statements hold.*

(i) For even steps $i = 2j$,

$$V(2j) = 2j + h,$$

and maximum in (17) is provided by a_1 only.

(ii) For odd steps $i = 2j-1$,

$$V(2j-1) < \frac{2\varepsilon(j-1) + (1+\varepsilon)h - 2}{\varepsilon}.$$

(iii) For odd steps $i = 2j+1$,

$$V(2j+1) = (1-\varepsilon)V(2j-1) + (1+\varepsilon)h + 2\varepsilon(j-1),$$

and maximum in (17) is provided by a_2 only.

For the proof see Lemma 3 of [20].

Therefore, for all odd values of i, the maximum in (17) is provided only by $a_2 \notin \mathbf{A}_*$. The conjecture (19) is not valid.

Now we look at this example from the viewpoint of the generalized model $\langle \mathbf{D}, \mathcal{R}, \mathcal{Q}\rangle$ described in Subsection 7.2: $M = 3$ and $\mathbf{S} = \{0, 1, 2\}$. The only trivial decision is $d^*(s) \equiv a_1$, which is certainly average optimal, as mentioned below Definition 2.

Lemma 10 *The only average optimal decision in the model $\langle \mathbf{D}, \mathcal{R}, \mathcal{Q}\rangle$ is the trivial decision d^*.*

The functions W^k, $k \geq 1$, coming from V, have the following form (see Lemma 9):

- If k is odd, then

$$\begin{aligned} W^k(0) &= \tilde{W}(3(k-1)) = V(3(k-1)) - 3(k-1) = h; \\ W^k(1) &= \tilde{W}((3(k-1)+1) = V(3(k-1)+1) - [3(k-1)+1] \\ &=: z(2j+1), \quad \text{where } j := 3(k-1)/2; \\ W^k(2) &= \tilde{W}((3(k-1)+2) = V(3(k-1)+2) - [3(k-1)+2] = h. \end{aligned}$$

- If k is even, then

$$\begin{array}{rcl} W^k(0) & = & \tilde{W}(3(k-1)) = V(3(k-1)) - 3(k-1) \\ & =: & z(2j+1), \quad \text{where } j := [3(k-1)-1]/2; \\ W^k(1) & = & \tilde{W}((3(k-1)+1) = V(3(k-1)+1) - [3(k-1)+1] = h \\ W^k(2) & = & \tilde{W}((3(k-1)+2) = V(3(k-1)+2) - [3(k-1)+2] \\ & =: & z(2j+1), \quad \text{where } j := [3(k-1)+1]/2. \end{array}$$

The maximum in (24) for $k \geq 1$ is provided only by the decisions, with some abuse of notations, represented as

$$\begin{array}{rcl} d_o & := & (a_2, a_1, a_2) \text{ if } k \text{ is odd, and} \\ d_e & := & (a_1, a_2, a_1) \text{ if } k \text{ is even.} \end{array} \tag{29}$$

According to Lemma 9(iii),

$$z(2j+1) + 2j + 1 = (1-\varepsilon)[z(2j-1) + 2j - 1] + (1+\varepsilon)h + 2\varepsilon(j-1), \quad j \geq 1,$$

i.e.,

$$z(2j+1) = (1-\varepsilon)z(2(j-1)+1) - 2 - \varepsilon + (1+\varepsilon)h,$$

and $z(1) = V(1) - 1 = h - 1$. It is obvious that the sequence $\{z(2j+1)\}_{j=0}^{\infty}$ increases, $z(2j+1) < \left[h + \frac{h-2}{\varepsilon} - 1\right] < h$ for all $j \geq 0$ (the rigorous proof is by induction), and $\lim_{j\to\infty} z(2j+1) = h + \frac{h-2}{\varepsilon} - 1$. Therefore, the sequence of vectors $\{W^k\}_{k=0}^{\infty}$ in $\mathbb{R}^M$ is uniformly bounded and $\lim_{k\to\infty} sp(W^{k+1} - W^k) = 2\left[1 - \frac{h-2}{\varepsilon}\right] > 0$ as expected because, if $\lim_{k\to\infty} sp(W^{k+1} - W^k) = 0$ then, by the Turnpike Theorem 1, the maximum in (24) would have been provided by the average optimal decision d^* for large enough k.

Condition 1 is violated because $\mathcal{Q}(d^*) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$, and the state 0 is here transient.

Condition 2 is violated because the matrix $\mathcal{Q}(d^*)$ is not aperiodic and, for each $J \geq 1$, for $d^1 = d^2 = \{d^*, d^*, \ldots, d^*\}$, we have

$$\eta(d^1, d^2) = \sum_{l=0}^{2} \min\{\mathcal{Q}_{1,l}^{(J,1)}, \mathcal{Q}_{2,l}^{(J,2)}\} = 0 \implies \gamma^J = 1.$$

Corollary 2 is not applicable because $p_3(a_1) = 0$. Proposition 2 and Corollary 3 are not applicable either.

The both decision d_o and d_e (29) are not average optimal (see Lemma 10). At the same time, the switching control strategies $\pi_1 := (d_e, d_0, d_e, \ldots)$ and $\pi_2 := (d_o, d_e, d_o, \ldots)$ are average optimal in the sense that

$$\lim_{T\to\infty} \frac{1}{T} \mathbb{E}_s^{\pi_{1,2}} \left[\sum_{t=1}^{T} \mathcal{R}_{S(t-1)}(d_t) \right] = 0, \qquad s \in \mathbf{S}. \tag{30}$$

(Cf (9) and see Lemma 11 below.) Here $\mathbb{E}_s^{\pi_{1,2}}$ is the mathematical expectation with respect to the strategical measure $\mathbb{P}_s^{\pi_{1,2}}$ on the trajectories $\omega = (s_0 = s, s_1, s_2, \ldots)$ in $\mathbf{S}$, which is defined in the standard for MDP way: $S(t, \omega) = s_t$ is the t-th component of ω (the argument ω is usually omitted), the process $S(t)$, $t = 0, 1, \ldots$ is Markov, and $\mathbb{P}_s^{\pi_{1,2}}(S(0) = s) = 1$;

$$\mathbb{P}_s^{\pi_{1,2}}(S(t) = j | S(t-1) = i) = \left\{ \begin{array}{ll} \mathcal{Q}_{i,j}(d_e) & \text{for } \pi_1, \text{ if } t \text{ is odd} \\ & \text{and for } \pi_2, \text{ if } t \text{ is even;} \\ \mathcal{Q}_{i,j}(d_o) & \text{for } \pi_1, \text{ if } t \text{ is even} \\ & \text{and for } \pi_2, \text{ if } t \text{ is odd.} \end{array} \right.$$

Similar rigorous constructions are presented in many monographs on MDP: see, e.g., [18, 21].

Lemma 11 *Both for π_1 and π_2, the equality (30) is valid.*

The situation when a combination of two (or more) non-optimal control strategies results in an optimal one, is called 'Parrondo's Paradox', see [19, §4.2.30].

8 Summary

We developed the turnpike theory for the undiscounted MDPs and showed that, without appropriate conditions, the turnpike property may be not valid. To the best of our knowledge, the existence of two turnpikes, as in Section 6 and Subsection 7.3, was never mentioned before. The presented theory is illustrated by an example of random walk, which seems to be important by its own.

It looks important to develop the turnpike theory for non-finite MDPs, discounted and undiscounted: publications on this topic are very limited [8, 14].

It is also desirable to understand when the turnpike property holds in the two-step (more generally, n-step) model: see the last paragraph of Section 3.

9 Appendix

Proof of Lemma 2. (a) For $k=2$ we obviously have $x_1 = 0 < \varepsilon - h = x_2 < \frac{\varepsilon - h}{\varepsilon}$ and, if the formulated inequalities are valid for some $k \geq 3$, then for $k+1$ we have

$$\begin{aligned} x_{k+1} &= (1-\varepsilon)x_k + \varepsilon - h = x_k + \varepsilon - h - \varepsilon x_k > x_k + \varepsilon - h - \varepsilon\left[1 - \frac{h}{\varepsilon}\right] = x_k; \\ x_{k+1} &< (1-\varepsilon)\left[1 - \frac{h}{\varepsilon}\right] + \varepsilon - h = 1 - \frac{h}{\varepsilon}. \end{aligned}$$

(b) The reasoning for the absorbing state 2 is trivial.

Clearly, $W^1(0) = 1$ and $W^1(1) = 0$ and

$$\begin{aligned} W^2(0) &= \max\{0+0,\ -h+\varepsilon = x_2,\ 1-2\} = x_2; \\ W^2(1) &= \max\{0+1,\ -h+1-\varepsilon,\ 0-2\} = 1: \end{aligned}$$

the expression for W^{k+1} in this lemma is valid at $k=1$.

Suppose it is valid for some $k-1 \geq 1$ and consider the case of k.

If k is even, then by induction

$$W^{k+1}(0) = \max\{0+1,\ -h+(1-\varepsilon)+\varepsilon x_k,\ 1-2k\} = 1$$

because $\varepsilon x_k - \varepsilon - h < \varepsilon x_k - \varepsilon + h < 0$, and the maximum is provided by a_1;

$$W^{k+1}(1) = \max\{0+x_k,\ -h+(1-\varepsilon)x_k+\varepsilon = x_{k+1},\ -2k\} = x_{k+1}$$

because $x_{k+1} > x_k$, and the maximum is provided by a_2.

If $k \geq 3$ is odd, then by induction

$$\begin{aligned} W^{k+1}(0) &= \max\{0+x_k,\ -h+(1-\varepsilon)x_k+\varepsilon = x_{k+1},\ 1-2k\} = x_{k+1}; \\ W^{k+1}(1) &= \max\{0+1,\ -h+(1-\varepsilon)+\varepsilon x_k,\ -2k\} = 1. \end{aligned}$$

The details are as those given above. □

Proof of Proposition 1. All the statements are well known for the standard model [12, 21]. In the general case, the proofs are word by word identical. We sketch them below.

(a-b) The both operator U^β with $\beta \in (0,1)$ and $V \to \mathcal{R}(d) + \beta\mathcal{Q}(d)V$ are contractions in the space $\mathbb{R}^M$ with the uniform norm. The maximum in (8) is provided by some $d^* \in \mathbf{D}$ according to Assumption 1.

The solutions $V^{\beta *}$ and $V^{\beta d}$ to the equations (8) and (13) can be built by iterations like (7) with $W^0 = 0$, leading to the sequences W^{*k} and W^k correspondingly. On each step $W^{*k} \ge W^k$; hence

$$V^{\beta *} = \lim_{k\to\infty} W^{*k} \ge \lim_{k\to\infty} W^k = V^{\beta d}.$$

(c) See [12, Thm.5.3] and also [13].

(d) See [21, Thm.A.7].

(e) For the existence of a Blackwell optimal decision, see [12, Thm.4.1] which proof remains unchanged in the general case. The same concerns the average optimality of each Blackwell optimal decision: see Thm.5.8(2) and Cor.5.3 of [12]. See also [21, Thm.8.4.5], where the similar statements were proved under the (unneeded) condition that MDP is unichain. □

Proof of Lemma 3. The reasoning is similar to the proof of Theorm 8.5.5 of [21]. Below, we provide the key statements.

According to Proposition 1(c), $\mathcal{Q}^*(\hat{d})\mathcal{Q}(\hat{d}) = \mathcal{Q}^*(\hat{d})$. Thus

$$\begin{aligned} \mathcal{Q}^*(\hat{d})\mathcal{R}(\hat{d}) &= \mathcal{Q}^*(\hat{d})\left[\mathcal{R}(\hat{d}) + \mathcal{Q}(\hat{d})W - W\right] = \mathcal{Q}^*(\hat{d})\left[U \circ W - W\right] \\ &\ge \min_{s\in\mathbf{S}} \{U \circ W(s) - W(s)\}\, e \end{aligned}$$

because $\mathcal{Q}^*(\hat{d})$ is a stochastic matrix.

For any average optimal decision d^*, which exists due to Proposition 1(e), we similarly have

$$\begin{aligned} \mathcal{Q}^*(\hat{d})\mathcal{R}(\hat{d}) &\le \mathcal{Q}^*(d^*)\mathcal{R}(d^*) = \mathcal{Q}^*(d^*)\left[\mathcal{R}(d^*) + \mathcal{Q}(d^*)W - W\right] \\ &\le \mathcal{Q}^*(d^*)\left[U \circ W - W\right] \le \max_{s\in\mathbf{S}} \{U \circ W(s) - W(s)\}\, e. \end{aligned}$$

□

Proof of Lemma 4. For the standard model Item (a) follows from Theorem 8.5.2 of [21]. For the general model $\langle \mathbf{D}, \mathcal{R}, \mathcal{Q}\rangle$, one can repeat the proof of Theorem 8.5.2 [21] word by word.

(b) Take N_1 such that, for all $n \ge N_1$,

$$sp(U^{nJ} \circ W^1 - U^{nJ} \circ W^0) = sp(W^{nJ+1} - W^{nJ}) \le \varepsilon;$$

take $N_2 \ge N_1$ such that, for all $n \ge N_2$,

$$sp(U^{nJ} \circ W^2 - U^{nJ} \circ W^1) = sp(W^{nJ+2} - W^{nJ+1}) \le \varepsilon;$$

and so on;
take $N_J \ge N_{J-1} \ge \ldots \ge N_1$ such that, for all $n \ge N_J$,

$$sp(U^{nJ} \circ W^J - U^{nJ} \circ W^{J-1}) = sp(W^{nJ+J} - W^{nJ+J-1}) \le \varepsilon.$$

Now, for $K = N_J$, $sp(W^{k+1} - W^k) \le \varepsilon$ for all $k \ge K$. □

Proof of Corollary 1. For any two non-negative $M \times M$ matrices P^1 and P^2 we write $P^1 \preceq P^2$ if, for each pair $(s,l) \in \mathbf{S}^2$, $P^1_{s,l} > 0 \Longrightarrow P^2_{s,l} > 0$. Since $\mathcal{Q}_{s,s}(d) > 0$ for all $s \in \mathbf{S}$, we have

$$P^1 \preceq P^1\mathcal{Q}(d) \quad \text{and} \tag{31}$$

$$\text{if } P^1 \preceq P^2, \quad \text{then} \quad P^1\mathcal{Q}(d) \preceq P^2\mathcal{Q}(d) \tag{32}$$

for all $d \in \mathbf{D}$ and for all non-negative $M \times M$ matrices P^1 and P^2.

For each $d \in \mathbf{D}$, there exists N_d such that $\mathcal{Q}^n(d) > 0$ for all $n \geq N_d$ [13, Thm.4.1.2]. Hence, since the set $\mathbf{D}$ is finite, for $N := \max_{d\in\mathbf{D}} N_d < \infty$, for all $d \in \mathbf{D}$, $\mathcal{Q}^n(d) > 0$ for all $n \geq N$. Let $J := (N-1)|\mathbf{D}| + 1$, where $|\mathbf{D}|$ is the total number of decisions in $\mathbf{D}$.

Fix an arbitrary sequence $\{d_1, d_2, \ldots, d_J\}$ of decisions from $\mathbf{D}$. Then at least one decision $\hat{d}$ appears at least N times in this list, say,

$$d_{j_1} = \hat{d},\ d_{j_2} = \hat{d}, \ldots, d_{j_N} = \hat{d} \ \text{ with } 0 < j_1 < j_2 < \ldots < j_N \leq J,$$

and $\mathcal{Q}^N(\hat{d}) > 0$.

According to (31),(32),

$$I \preceq \mathcal{Q}(d_1) \preceq \mathcal{Q}(d_1)\mathcal{Q}(d_2) \preceq \ldots \preceq \mathcal{Q}(d_1)\mathcal{Q}(d_2)\ldots\mathcal{Q}(d_J) = \mathcal{Q}^{(J)},$$

where I is the identical $M \times M$ matrix.

According to (32),

$$\mathcal{Q}(\hat{d}) = \mathcal{Q}(d_{j_1}) \preceq \mathcal{Q}(d_1)\mathcal{Q}(d_2)\ldots P(d_{j_1})$$

because $I \preceq \mathcal{Q}(d_1)\mathcal{Q}(d_2)\ldots P(d_{j_1-1})$. Similarly

$$\mathcal{Q}^2(\hat{d}) = \mathcal{Q}(\hat{d})\mathcal{Q}(\hat{d}) \preceq \mathcal{Q}(d_1)\mathcal{Q}(d_2)\ldots P(d_{j_2})$$

because $\mathcal{Q}(\hat{d}) \preceq \mathcal{Q}(d_1)\mathcal{Q}(d_2)\ldots P(d_{j_1}) \preceq \mathcal{Q}(d_1)\mathcal{Q}(d_2)\ldots P(d_{j_2-1})$.
And so on:

$$\mathcal{Q}^N(\hat{d}) = \mathcal{Q}^{N-1}(\hat{d})\mathcal{Q}(\hat{d}) \preceq \mathcal{Q}(d_1)\mathcal{Q}(d_2)\ldots P(d_{j_N})$$

because $\mathcal{Q}^{N-1}(\hat{d}) \preceq \mathcal{Q}(d_1)\mathcal{Q}(d_2)\ldots P(d_{j_{N-1}}) \preceq \mathcal{Q}(d_1)\mathcal{Q}(d_2)\ldots P(d_{j_N-1})$.

Therefore, $\mathcal{Q}^N(\hat{d}) \preceq \mathcal{Q}^{(J)}$ meaning that $\mathcal{Q}^{(J)} > 0$, and $\gamma^J < 1$ by Lemma 5.

□

Proof of Lemma 6. All the reasoning is similar to the proof of Corollary 9.4.6 of [21]. The milestones are provided below.

(i) For each decision $d \in \mathbf{D}$, let

$$g(d) := \mathcal{Q}^*(d)\mathcal{R}(d) \ \text{ and } h(d) := D(d)\mathcal{R}(d)$$

be the so called gain and bias of the decision d. Then

$$g(d) = \mathcal{Q}(d)g(d) \ \text{ and } g(d) + (I - \mathcal{Q}(d))h(d) = \mathcal{R}(d).$$

(See [21, Thm.8.2.6].)

(ii) There exists a solution $(g, Y) \in \mathbb{R}^M \times \mathbb{R}^M$ to the (so called 'modified') optimality equation

$$g = \max_{d\in\mathbf{D}}\{\mathcal{Q}(d)g\}; \qquad g + Y = \max_{d\in\mathbf{D}}\{\mathcal{R}(d) + \mathcal{Q}(d)Y\}, \tag{33}$$

for which necessarily $g = g^* := \max_{d\in\mathbf{D}} \mathcal{Q}^*(d)\mathcal{R}(d)$ is the maximal gain [12, Cor.5.4], [21, Thm.9.1.2, Cor.9.1.5]. If a decision d^* provides the both maxima, then d^* is average optimal [21, Thm.9.1.7].

(iii) $\lim_{k\to\infty} \frac{1}{k} W^k = g^*$.
See [21, Thm.9.4.1] and also [12, Lemma 5.5].

(iv) Suppose, for every average optimal decision $d^* \in \mathbf{D}^*$, the matrix $\mathcal{Q}(d^*)$ is aperiodic. Then there exists the limit

$$\lim_{k\to\infty} [W^k - kg^*]$$

[21, Thm.9.4.4].

(v) Therefore,

$$\mathbf{0} = \lim_{k\to\infty} \left[[W^{k+1} - (k+1)g^*] - [W^k - kg^*] \right] = \lim_{k\to\infty} \left[(W^{k+1} - W^k) - g^* \right],$$

where $\mathbf{0}$ is the zero vector in $\mathbb{R}^M$.

(vi) Under the assumptions of Lemma 6,

$$\lim_{k\to\infty} (W^{k+1} - W^k) = G^* e \Longrightarrow \lim_{k\to\infty} \ sp(W^{k+1} - W^k) = 0.$$

All the statements (i)-(vi) formulated above were proved in [12, 21] for the standard model; all the steps of the proofs remain valid for the general case. □

Proof of Lemma 8. Suppose $d^* \in \mathbf{D}^*$, that is,

$$\mathcal{Q}^*(d^*)\mathcal{R}(d^*) = \max_{d\in\mathbf{D}} \mathcal{Q}^*(d)\mathcal{R}(d) = \mathbf{0}.$$

Matrix $\mathcal{Q}^*(d^*)$ in the block representation ('standard form') has only the blocks corresponding to the closed communicating classes, and those blocks are strictly positive [12, §5.3.1]. Thus, if $\mathcal{R}(d^*) \neq \mathbf{0}$ then the non-positive vector $\mathcal{Q}^*(d^*)\mathcal{R}(d^*)$ contains negative elements, and $d^* \notin \mathbf{D}^*$. The obtained contradiction completes the proof. □

Proof of Corollary 2. Let us show that, for each $d \in \mathbf{D}$, for any two states $i, j \in \mathbf{S}$, $\mathcal{Q}_{i,j}(d) > 0$ provided $P_{i,j}(d(i)) > 0$.

If $j \geq i$ then this statement follows directly from the definition of the matrix $\mathcal{Q}(d)$:

$$\mathcal{Q}_{i,j}(d) \geq P_{i,j}(d(i)) > 0.$$

Suppose $j < i$. Then, again using the definition of he matrix $\mathcal{Q}(d)$, we have

$$\mathcal{Q}_{i,j}(d) \geq P_{i,j}(d(i))\mathcal{Q}_{j,j}(d).$$

Since $\mathcal{Q}_{j,j}(d) \geq P_{j,j}(d(j)) = p_M(d(j)) > 0$, we obtain the required inequality $\mathcal{Q}_{i,j}(d) > 0$, if $P_{i,j}(d(i)) > 0$.

Now, for any two states $i, j \in \mathbf{S}$, for the path $i_0 = i \to i_1 \to \ldots \to i_N = j$ in $\mathbf{S}$, we have

$$\mathcal{Q}_{i_0,i_1}(d)\mathcal{Q}_{i_1,i_2}(d)\ldots\mathcal{Q}_{i_{N-1},i_N}(d) > 0$$

for each $d \in \mathbf{D}$. To put it different, for each $d \in \mathbf{D}$, the matrix $\mathcal{Q}(d)$ corresponds to the irreducible Makov chain which is certainly aperiodic because $\mathcal{Q}_{i,i}(d) \geq P_{i,i}(d(i)) > 0$ for all $i \in \mathbf{S}$.

Since Conditions 1 and 2 are satisfied, the statement of Corollary follows from Theorem 3. □

Proof of Lemma 10. One should consider all the different decisions $d : \mathbf{S} \to \mathbf{A}$. Since the reasoning is similar in all the cases, below we investigate only the decision $d(0) = a_1$, $d(1) = a_1$, $d(2) = a_2$. Here

$$\mathcal{R}(d) = (0, 0, h - (2+\varepsilon))^T; \qquad \mathcal{Q}(d) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1-\varepsilon & \varepsilon \end{pmatrix}.$$

Recall that the model $\langle \mathbf{D}, \mathcal{R}, \mathcal{Q}\rangle$ is as in the version (ii) described in Section 3; $r_s(a) := L^a\left(\frac{R^a}{L^a} - c_*\right)$ and the matrix $P(a)$ is given by the equation (22). Remember also that $h-(2+\varepsilon) < 0$. Clearly, the states 1 and 2 form the closed communicating class, the state 0 is transient, and the components $\mathcal{Q}^*_{0,2}(d), \mathcal{Q}^*_{1,2}(d), \mathcal{Q}^*_{2,2}(d)$ are strictly positive, so that $\mathcal{Q}^*(d)\mathcal{R}(d) < 0$, and the decision d is not average optimal.

For all other non-trivial decisions d, one can similarly show that the vector $\mathcal{Q}^*(d)\mathcal{R}(d)$ contains strictly negative components. □

Proof of Lemma 11. The reasoning is similar for π_1 and π_2, so we consider only $\pi_1 = (d_e, d_0, d_e, \ldots)$. Obviously, it is sufficient to show that the sequence of vectors in $\mathbb{R}^M$

$$R^T := \left\{\mathbb{E}^{\pi_1}_s\left[\sum_{t=3}^{T}\mathcal{R}_{S(t-1)}(d_t)\right],\ s\in\mathbf{S}\right\},\quad T\geq 3,$$

is uniformly bounded.

Suppose an odd value of $T\geq 3$ is fixed and consider the sequence of vectors $\{\tilde{R}^l_i,\ i\in\mathbf{S}\}_{l=2}^{T}$ in $\mathbb{R}^M$ defined as follows:

$$\tilde{R}^l_i := \mathbb{E}^{\pi_1}_s\left[\sum_{t=T-l+3}^{T}\mathcal{R}_{S(t-1)}(d_t) + W^2(S(T))|S(T-l+2)=i\right].$$

These vectors are s-independent due to the Markov property of the process $\{S(t),\ t=0,1,\ldots\}$. Since the maximum in (24) is provided by d_e (d_o) if k is even (odd),

$$\begin{array}{rcl}\tilde{R}^2 &=& W^2;\\ \tilde{R}^3 &=& \mathcal{R}(d_e)+\mathcal{Q}(d_e)\tilde{R}^2 = W^3;\\ \tilde{R}^4 &=& \mathcal{R}(d_o)+\mathcal{Q}(d_o)\tilde{R}^3 = W^4;\quad \ldots,\\ \tilde{R}^T &=& \left\{\mathbb{E}^{\pi_1}_s\left[\sum_{t=3}^{T}\mathcal{R}_{S(t-1)}(d_t)+W^2(S(T))|S(2)=i\right],\ i\in\mathbf{S}\right\}\\ &=& \mathcal{R}(d_e)+\mathcal{Q}(d_e)\tilde{R}^{T-1} = W^T.\end{array}$$

The sequence $\{W^T\}_{T=0}^{\infty}$ is uniformly bounded by Lemma 7; so the sequence of vectors

$$\begin{array}{rcl}R^T &=& \left\{\mathbb{E}^{\pi_1}_s\left[\mathbb{E}^{\pi_1}_s\left[\sum_{t=3}^{T}\mathcal{R}_{S(t-1)}(d_t)+W^2(S(T))|S(2)\right]\right]\right.\\ && \left.-\,\mathbb{E}^{\pi_1}_s\left[\mathbb{E}^{\pi_1}_s\left[W^2(S(T))|S(2)\right]\right],\ s\in\mathbf{S}\right\}\\ &=& \left\{\mathbb{E}^{\pi_1}_s\left[W^T(S(2))\right]-\mathbb{E}^{\pi_1}_s\left[W^2(S(T))\right],\ s\in\mathbf{S}\right\},\quad T=3,5,7,\ldots\end{array}$$

is uniformly bounded.

For even values of T, the sequence

$$R^T := \left\{\mathbb{E}^{\pi_1}_s\left[\sum_{t=3}^{T-1}\mathcal{R}_{S(t-1)}(d_t)\right]+\mathbb{E}^{\pi_1}_s\left[\mathcal{R}_{S(T-1)}(d_T)\right],\ s\in\mathbf{S}\right\},\quad T=4,6,8,\ldots$$

is also uniformly bounded, and the proof is completed. □

References

[1] Chen, K. and Ross, S.M.: An adaptive stochastic knapsack problem. *European J. of Oper. Res.*, **239** (2014) 625–635.

[2] Dean, B.C., Goemans, M.X. and Vondrak, J.: Approximating the stochastic knapsack problem: the benefit of adaptivity. *Math. Oper. Res.*, **33** (2008) 945–964.

[3] Denardo, E.V. and Rothblum, U.G.: A turnpike theorem for a risk-sensitive Markov decision process with stopping. *SIAM J. Control Optim.*, **45** (2006) 414–431.

[4] Dorfman, R., Samuelson, P.A. and Solow, R.M.: *Linear Programming and Economic Analysis.* McGraw-Hill, NY, 1958.

[5] Federgruen, A. and Schweitzer, P.J.: Discounted and undiscounted value-iteration in Markov decision problems: a survey. In: Puterman, M.L. (ed) *Dynamic Programming and Its Applications*, pp. 23–52. Academic Press, NY, 1978.

[6] Guasoni, P., Kardaras, C., Robertson, S. and Xing, H.: Abstract, classic, and explicit turnpikes. *Financ. Stoch.*, **18** (2014) 75–114.

[7] Haurie, A. and Van Delft: Turnpike properties for a class of piecewise deterministic systems arising in manufacturing flow control. *Ann. Oper. Res.*, **29** (1991) 351–373.

[8] Hernandez-Lerma, O. and Lasserre, J.B.: A forecast horizon and a stopping rule for general Markov decision processes. *J. Math. Anal. Appl.*, **132** (1988) 388–400.

[9] Hinderer, K. and Waldmann, K.-H.: Algorithms for countable state Markov decision models with an absorbing set. *SIAM J. Control Optim.*, **43** (2005) 2109–2131.

[10] Iida, T. and Mori, M.: Markov decision processes with random horizon. *J .Oper. Res. Soc. Jpn.*, **39** (1996) 592–603.

[11] Jacobson, M., Shimkin, N. and Shwartz, A.: Markov decision processes with slow scale periodic decisions. *Math. Oper. Res.*, **28** (2003) 777–800.

[12] Kallenberg, L.C.M.: *Markov Decision Processes.* Lecture Notes, University of Leiden, The Netherlands, 2010.

[13] Kemeni, J.G. and Snell, J.L.: *Finite Markov Chains.* Van Nostrand Co., Princeton, 1960.

[14] Kolokoltsov, V.N.: Turnpikes and infinite extremals in Markov decision processes. *Mat. Zametki*, **46** (1989) 118–120 (in Russian).

[15] Lewis, M.E. and Paul, A.: Uniform turnpike theorems for finite Markov decision processes. *Math. Oper. Res.*, **44** (2019) 1145–1160.

[16] Morton, T.E.: Decision horizons in discrete time undiscounted Makov renewal programming. *IEEE Trans. Systems Man Cybernet.*, **SMC-4** (1974) 392–394.

[17] Odoni, A.R.: On finding the maximal gain for Markov decision processes. *Operations Research*, **17** (1969) 857–860.

[18] Piunovskiy, A.: *Optimal Control of Random Sequences in Problems with Constraints.* Kluwer, Dordrecht, 1997.

[19] Piunovskiy, A.: *Examples in Markov Decision Processes.* Imperial College Press, London, Vol.2, 2013.

[20] Piunovskiy, A.: Controlled random walk: conjecture and counter-example. In: Piunovskiy, A. and Zhang, Y. (eds) *Modern Trends in Controlled Stochastic Processes: Theory and Applications. Vol.III*, pp. 38–56. Springer, Cham, 2021.

[21] Puterman, M.: *Markov Decision Processes.* John Wiley & sons, NY, 1994.

[22] Sethi, S.P., Yan, H., Zhang, H. and Zhang, Q.: Turnpike set analysis in stochastic manufacturing systems with long-run average cost. In: Menaldi, J.L., Rofman, E. and Sulem, A. (eds) *Optimal Control and Partial Differential Equations*, pp. 414–423. IOS Press, Amsterdam, 2001.

[23] Shapiro, J.F.: Turnpike planning horizons for a markovian decision model. *Manage. Sci.*, **14** (1968) 292–300.

[24] Zaslavski, A.J.: *Turnpike Properties in the Calculus of Variations and Optimal Control.* Springer, New York, 2006.

[25] Zaslavski, A.J.: *Structure of Approximate Solutions of Optimal Control Prolems.* Springer, New York, 2013.

[26] Zaslavski, A.J.: *Turnpike Theory for the Robinson–Solow–Srinivasan Model.* Springer, Switzerland, 2021.