Optimal policies and restless bandits for making costly observations

Christopher R. Dance, NAVER LABS Europe

Workshop on Restless Bandits, Grenoble, November, 2023

Outline

problems		results		
observing a single time series	⟵	optimality of threshold policies	⟵	why?
controlling a single time series	⟵	optimality of threshold policies		
restless bandit for multiple time series	⟵	existence of Whittle index		

Based on: Dance and Silander, Optimal Policies for Observing Time Series and Related Restless Bandit Problems, *Journal of Machine Learning Research*, Vol. 20, No. 35, pp. 1-93, April 2019.

Problem 1: observing a single time series

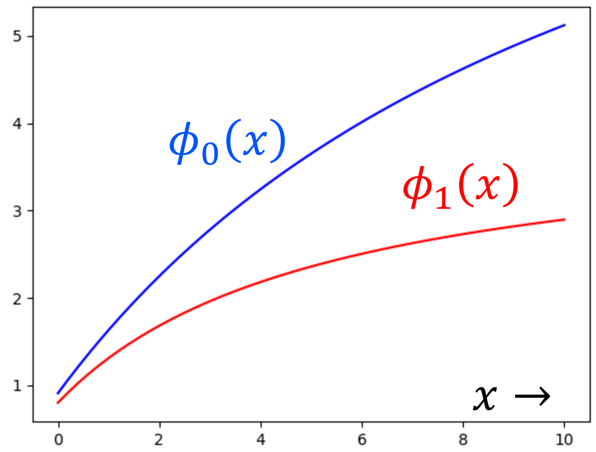
Discrete-time scalar normally distributed time series $X_0, X_1, \ldots$

$$X_{t+1} = r\, X_t + N(0,1)$$

Measurement action $a_t \in \{0,1\}$ results in measurement $Y_t \sim N\left(X_t, \frac{1}{\theta_{a_t}}\right)$

Posterior variance $x_t := \mathrm{var}(X_t \mid a_0, Y_0, \ldots, a_{t-1}, Y_{t-1})$ has Kalman filter update:

$$x_{t+1} = \frac{r^2 x_t + 1}{\theta_{a_t}(r^2 x_t + 1) + 1} =: \phi_{a_t}(x_t)$$



Uninformative observation $\theta_a = 0 \;\Rightarrow\; \phi_a(x) = r^2 x + 1$

Assume action $a = 0$ is less precise ($\theta_0 < \theta_1$) but action $a = 1$ costs $\lambda > 0$.

Problem. *When should we make costly-but-precise measurements of the time series to achieve a good trade-off between our uncertainty about the time series and the cost of precise observations?*

Problem 1: observing a single time series

Infinite horizon discounted Markov decision problem

state $x_t \in \mathbb{R}_{\geq 0}$ is the posterior variance

action $a_t = 0$ for a poor observation, $a_t = 1$ for a good observation

cost $x_t + \lambda a_t$

transition $x_{t+1} = \frac{r^2 x_t + 1}{\theta_{a_t}(r^2 x_t + 1) + 1} =: \phi_{a_t}(x_t)$ is Kalman filter variance update

discount $\beta \in (0,1)$

Dynamic programming equation for value function $V: \mathbb{R}_{\geq 0} \to \mathbb{R}$

$$V(x) = \min_{a \in \{0,1\}} \left\{ x + \lambda\, a + \beta\, V\big(\phi_a(x)\big) \right\}$$

Problem 2: controlling a single time series (Meier et al., 1967)

Discrete-time scalar linear quadratic Gaussian (LQG) control problem with costly measurements

$$X_{t+1} = r\,X_t + N(0,1) + \boldsymbol{U_t}$$

Problem. Select control $\boldsymbol{U_t} \in \mathbb{R}$ and measurement action $a_t \in \{0,1\}$ to minimize discounted sum of

$\mathbb{E}$ (quadratic penalty on X_t) + $\mathbb{E}$ (quadratic penalty on U_t) + $\mathbb{E}$ (pay λ each time we take action $a_t = 1$)

Fact. Problem separates into two independent parts:

- determining U_t given posterior mean for X_t
- determining measurement actions a_t

Dynamic programming equation for value function $V\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$

$$V(x) = \min_{a \in \{0,1\}} \left\{ \alpha\, x + \lambda\, a + \beta\, V\big(\phi_a(x)\big) \right\} \quad \text{for some constant } \alpha > 0$$

Problem 3: observing or controlling multiple time series (Villar, 2012)

Restless bandit problem for n scalar time series $\leftrightarrow$ parking occupancy statistics of n street segments

state	posterior variances for each of the n street segments
action	select $m < n$ street segments to observe with m cameras
transitions	Kalman filter variance update for each street
cost	sum of the n posterior variances

Problem. Does each project of this restless bandit have a well-defined Whittle index?

Main results: optimality of threshold policies (Dance and Silander, 2019)

Problem $\boldsymbol{P_\lambda}$ $\quad V(x) = \min_{a \in \{0,1\}} \left\{ x + \lambda\, a + \beta\, V\big(\phi_a(x)\big) \right\}$ where $\quad \phi_a(x) = \frac{r^2 x + 1}{\theta_a (r^2 x + 1) + 1}$

Thm. *Let multiplier $r \in [0,1]$, precisions $0 \le \theta_0 < \theta_1$ and discount $\beta \in (0,1)$.*

Then for some threshold $s \in [-\infty, \infty]$ an optimal policy for problem P_λ is to take
action $a = 1$ if $x \ge s$
action $a = 0$ if $x \le s$.

Remark. This theorem also holds for a wide range of cost functions $C\colon \mathbb{R}_{\ge 0} \to \mathbb{R}$

$$V(x) = \min_{a \in \{0,1\}} \left\{ C(x) + \lambda\, a + \beta\, V\big(\phi_a(x)\big) \right\}$$

including $C(x) = x^p$ for all $p > 0$ and $C(x) = -x^p$ for $p \in [-1, 0)$.

Cor. *Let $r \in [0,1]$, $0 \le \theta_0 < \theta_1$ and $\beta \in (0,1)$. Then a threshold policy is also optimal for making observations in the LQG problem with costly measurements.*

Main results: Whittle index

Problem $\boldsymbol{P_\lambda}$ $\quad V(x;\lambda) = \min_{a\in\{0,1\}} \{x + \lambda\, a + \beta\, V(\phi_a(x);\lambda)\}$

Def. *The **Whittle index*** of state x in problem P_λ is a real number $\lambda^*(x)$ such that

- *action* $a = 1$ *is optimal in state* x *if and only if* $\lambda \le \lambda^*(x)$
- *action* $a = 0$ *is optimal in state* x *if and only if* $\lambda \ge \lambda^*(x)$

Notation. Let $\mathbb{E}^{\boldsymbol{x,a,s}}$ denote the expectation for: start in x, take action a, follow s-threshold policy $a_t = 1_{x_t > s}$

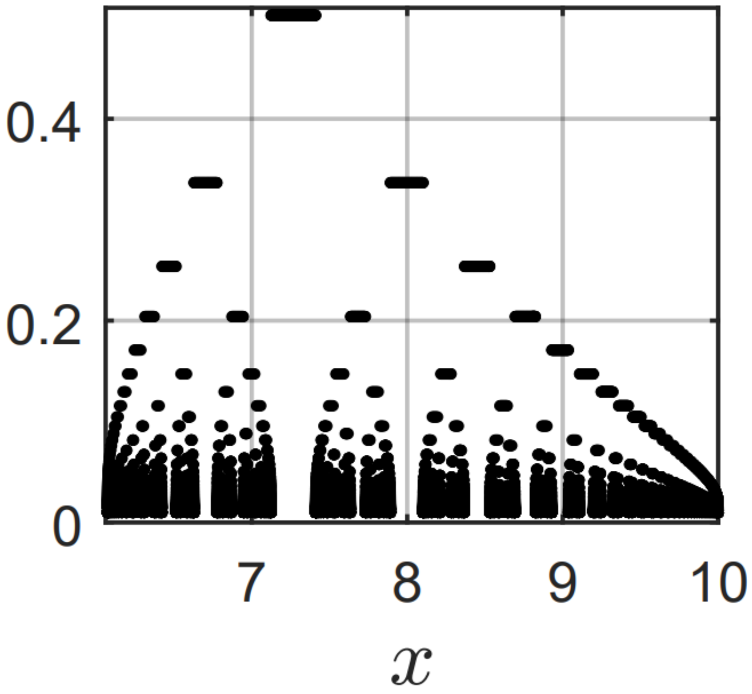
Thm. *Let multiplier* $r \in [0,1]$, *precisions* $0 \le \theta_0 < \theta_1$ *and discount* $\beta \in (0,1)$. *Then the Whittle index for the family of problems* P_λ *exists and equals*

$$\lambda^*(x) = \frac{\sum_{t=0}^{\infty} \beta^t (\mathbb{E}^{x,0,x} x_t - \mathbb{E}^{x,1,x} x_t)}{\sum_{t=0}^{\infty} \beta^t (\mathbb{E}^{x,1,x} a_t - \mathbb{E}^{x,0,x} a_t)}.$$

NAVER LABS EUROPE

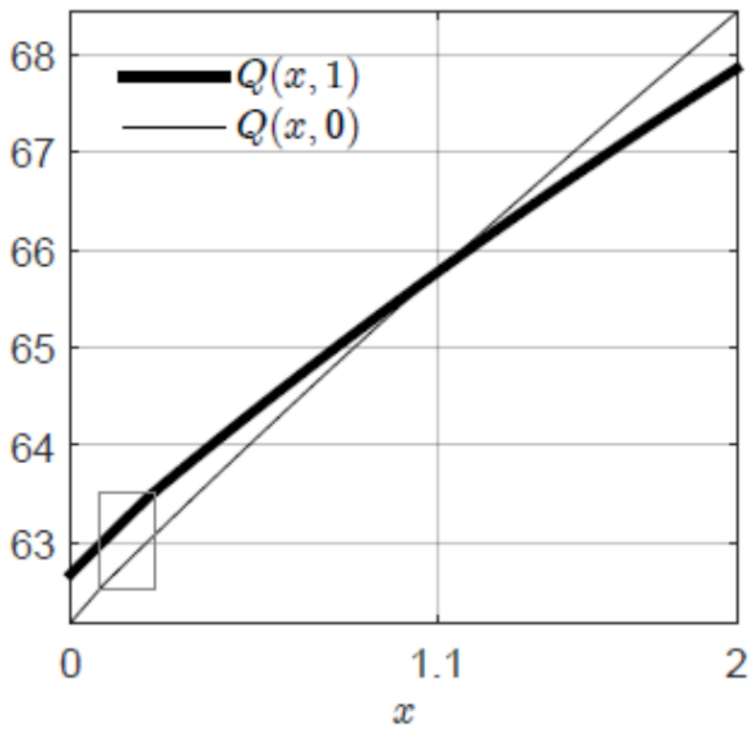
Plot of the denominator of the index

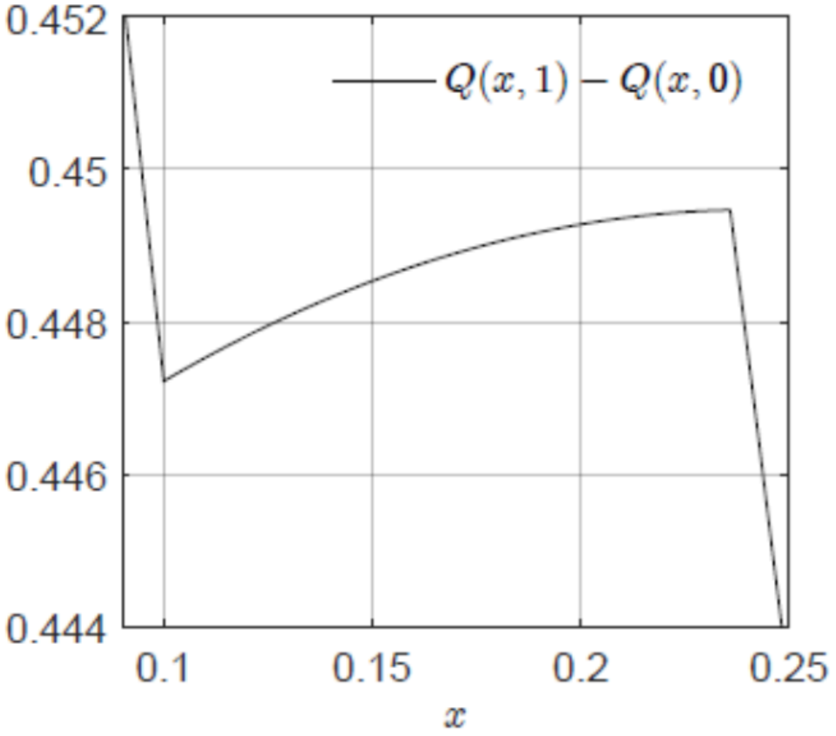
$$\mathbb{E}^{x,1,x}\sum_{t=0}^{\infty}\beta^t\, a_t - \mathbb{E}^{x,0,x}\sum_{t=0}^{\infty}\beta^t\, a_t$$



Why are threshold policies optimal?

Is $Q(x,1) - Q(x,0)$ monotone? (Serfozo, 1976)





Outline of proof

optimality of threshold policies

$g(x,s) > 0$

$m^*(x) \nearrow$

$$m^* = \frac{\mathrm{d}F}{\mathrm{d}G}$$

Niño-Mora (2019)

ϕ_0 and ϕ_1 are $\nearrow$ and
$\phi_1 \circ \phi_0(x) < \phi_0 \circ \phi_1(x)$

$\boldsymbol{p}$ is a palindrome

\# distinct prefixes of length n
of action sequences is $O(n^3)$
(Mignosi, 1991)

threshold policy for ϕ_a
$\Rightarrow$
action sequence given by
Christoffel words $\mathbf{0}\boldsymbol{p}\mathbf{1}$
and ***Sturmian words***

Outline of proof

optimality of threshold policies

$g(x,s) > 0$

$m^*(x) \nearrow$

$$m^* = \frac{\mathrm{d}F}{\mathrm{d}G}$$

Niño-Mora (2019)

marginal resource

$$g(x,s) \coloneqq \sum_{t=0}^{\infty} \beta^t (\mathbb{E}^{x,1,s} a_t - \mathbb{E}^{x,0,s} a_t)$$

marginal productivity index

$$m^*(x) \coloneqq \frac{\sum_{t=0}^{\infty} \beta^t (\mathbb{E}^{x,0,x} x_t - \mathbb{E}^{x,1,x} x_t)}{\sum_{t=0}^{\infty} \beta^t (\mathbb{E}^{x,1,x} a_t - \mathbb{E}^{x,0,x} a_t)}$$

reward metric

$$F(x,s) \coloneqq \sum_{t=0}^{\infty} \beta^t \, \mathbb{E}^{x,1_{x>s},s} \, (-x_t)$$

resource metric

$$G(x,s) \coloneqq \sum_{t=0}^{\infty} \beta^t \, \mathbb{E}^{x,1_{x>s},s} a_t$$

Outline of proof

optimality of threshold policies

$g(x,s) > 0$

$m^*(x) \nearrow$

$m^* = \frac{\mathrm{d}F}{\mathrm{d}G}$

Niño-Mora (2019)

ϕ_0 and ϕ_1 are $\nearrow$ and $\phi_1 \circ \phi_0(x) < \phi_0 \circ \phi_1(x)$

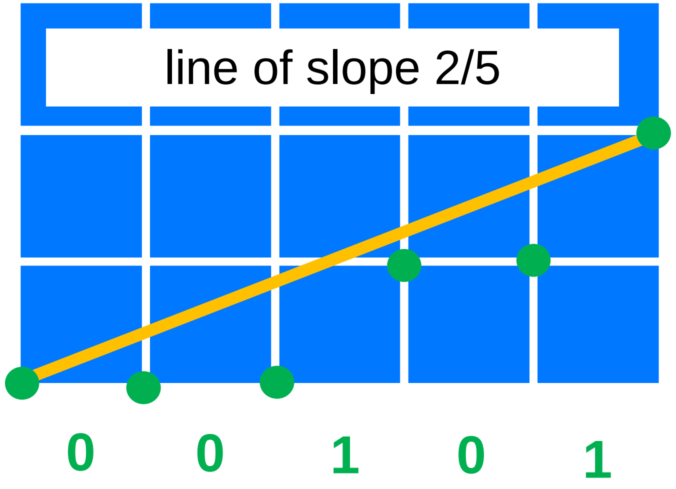
$\boldsymbol{p}$ is a palindrome

distinct prefixes of length n of action sequences is $O(n^3)$ (Mignosi, 1991)

threshold policy for ϕ_a
$\Rightarrow$
action sequence given by ***Christoffel words*** $\mathbf{0}\boldsymbol{p}\mathbf{1}$ and ***Sturmian words***

NAVER LABS EUROPE

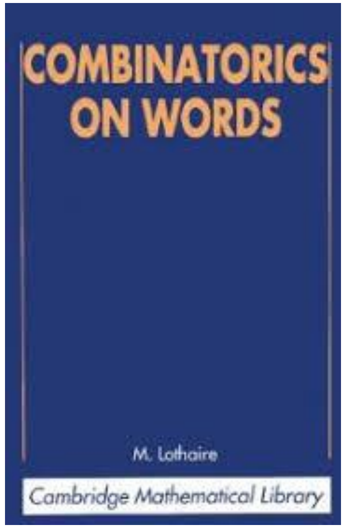
Action sequences resulting from threshold policies

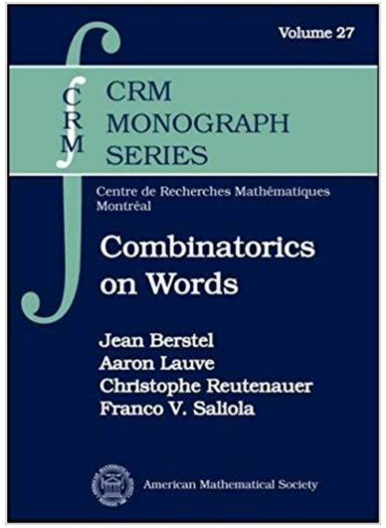


Def. *The* Christoffel word *of slope* a, *which is a rational number in* $[0,1]$, *is*

$$w_n = \lfloor (n+1)\, a \rfloor - \lfloor n\, a \rfloor$$

for $n = 0, 1, \ldots, \mathrm{denom}(a) - 1$





Ex. of Christoffel words: 0, 1, 01, 001, 011, 00101, …

~~0~~010~~1~~ → 010 is a palindrome

Prop. If $0p1$ *is a Christoffel word then* p *is a palindrome.*

conjugates

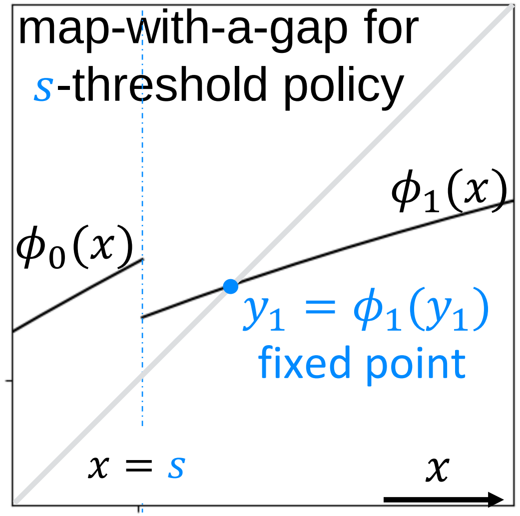
conjugates	sort	sorted
00101		$\boxed{0\,0\,1\,0\,1}$
01010		0 1 0 0 1
10100	→	0 1 0 1 0
01001		1 0 0 1 0
10010		$\boxed{1\,0\,1\,0\,0}$

Pirillo (1999). *A word* $0p1$ *is Christoffel if and only if it is conjugate to* $1p0$.

$\Rightarrow 01p$ and $10p$ are conjugates

NAVER LABS EUROPE

Action sequences resulting from threshold policies



Assp A. *For some real interval $\mathcal{I}$, maps $\phi_0 : \mathcal{I} \to \mathcal{I}$ and $\phi_1 : \mathcal{I} \to \mathcal{I}$ are*

- *increasing*
- *contractive* (i.e., $|\phi_a(x) - \phi_a(y)| < |x - y|,\ x, y \in \mathcal{I},\ x \neq y$)
- *have fixed points $y_1 < y_0$ in $\mathcal{I}$.*

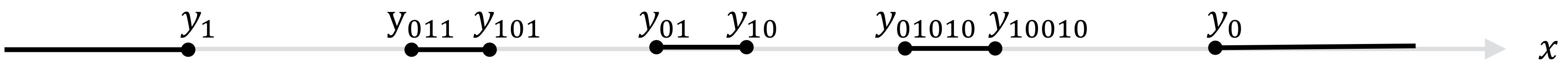
Def. For word $\mathrm{w} = w_1 \dots w_n$, define the composition $\phi_w \coloneqq \phi_{w_n} \circ \dots \circ \phi_{w_1}$ and its fixed point $y_w = \phi_w(y_w)$.

Thm. *Let Assp. A hold. Let the initial state be x. Then the action sequence under the x-threshold policy is*

01^∞ *if and only if* $x \leq y_1$

$(01p)^\infty$ *if and only if* $y_{01p} \leq x \leq y_{10p}$ *for any Christoffel word* $0p1$

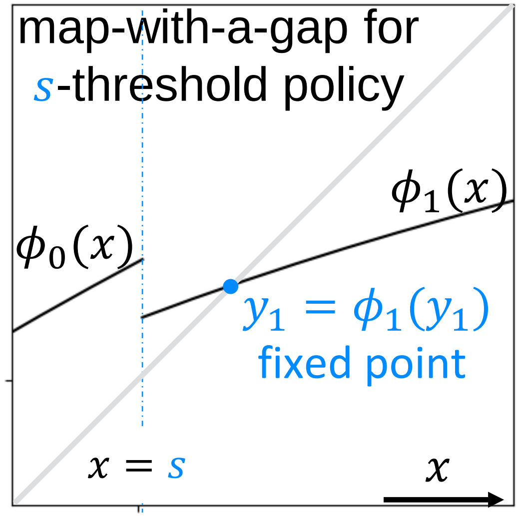
0^∞ *if and only if* $x \geq y_0$.



01^∞	…	$(011)^\infty$	…	$(01)^\infty$	…	$(01010)^\infty$	…	0^∞	→	action sequence
1	…	$\frac{2}{3}$	…	$\frac{1}{2}$	…	$\frac{2}{5}$	…	0	→	slope

NAVER LABS EUROPE

Action sequences resulting from threshold policies

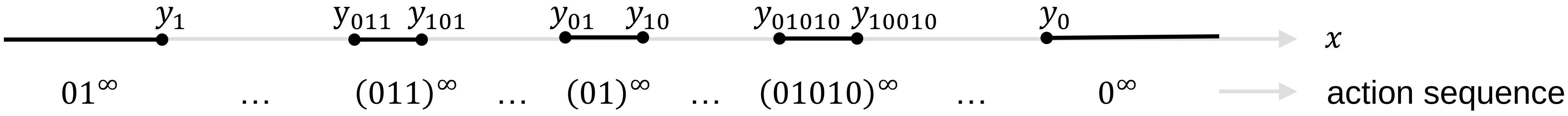


Assp A. *For some real interval $\mathcal{I}$, maps $\phi_0 : \mathcal{I} \to \mathcal{I}$ and $\phi_1 : \mathcal{I} \to \mathcal{I}$ are*
- *increasing*
- *contractive* *(i.e., $|\phi_a(x) - \phi_a(y)| < |x - y|,\ x, y \in \mathcal{I},\ x \neq y$)*
- *have fixed points $y_1 < y_0$ in $\mathcal{I}$.*

Def. For word $\mathrm{w} = w_1 \dots w_n$, define the composition $\phi_w \coloneqq \phi_{w_n} \circ \cdots \circ \phi_{w_1}$ and its fixed point $y_w = \phi_w(y_w)$.

Thm. *Let Assp. A hold. Let the initial state be x. Then the action sequence under the x-threshold policy is*

01^∞ *if and only if* $x \leq y_1$
$(01p)^\infty$ *if and only if* $y_{01p} \leq x \leq y_{10p}$ *for any Christoffel word* $0p1$
0^∞ *if and only if* $x \geq y_0$.



Rmk. This result was previously only partially known.
- Rajpathak *et al.* (2012) - only for *linear* maps.
- Kozyakin (2003) - for nonlinear maps but unclear dependence on x.

Thank you!