Research Article

Evgueni Gordienko* and Juan Ruiz de Chavez

Stability estimation of some Markov controlled processes

https://doi.org/10.1515/math-2022-0514
received February 11, 2022; accepted September 21, 2022

Abstract: We consider a discrete-time Markov controlled process endowed with the expected total discounted reward. We assume that the distribution of the underlying random vectors is unknown and that it is approximated by an appropriate known distribution. We found upper bounds of a decrease in reward when the policy, optimal for the approximating process, is applied to control the original process.

Keywords: optimal control policy, stability inequality, the total variation and the Dudley metrics

MSC 2020: 90B05, 90C31, 90C40, 93E20

1 Introduction

In the theory of discrete-time Markov processes, the term "stability" is used in various meanings. First and foremost, uncontrolled processes, this refers to some recurrent or ergodic properties of the processes (see, e.g., [1]).

Quite a long time ago, this concept moved into the field of controlled processes, particularly, in the context of adaptive control. (Among the huge number of references, we indicate only a couple of fairly recent ones [2,3].)

The second widely used meaning of the word "stability" is close to "continuity." Speaking of the quantitative approach to such continuity under perturbations of certain parameters, the deviations of some basic characteristics of the Markov processes (such as the limiting distribution) are estimated.

Using probability metrics, the methods of quantitative continuity of uncontrolled processes have been developed, for instance, in the works [4–6].

The quantitative assessment of the stability (or "continuity") of optimal control of a Markov process has its own peculiarities. Here, the policy that is optimal for a certain "approximating process" is used to control the original ("real") process. The underlying probability distributions of the latter are unknown and are often evaluated by statistical procedures. Such estimation leads to what we have designated as the "approximating controlled process."

The problem is posed as finding the upper bounds of *the stability index*, which is defined in (2.7) in Section 2, and it expresses the decrease in the given performance index (compared to the application of the optimal for the original process control). This problem was probably first considered in [7,8]. Since then, the authors just mentioned and others have been solving this problem for various classes of discrete-time Markov controlled processes and for different performance indexes (optimization criteria).

*** Corresponding author: Evgueni Gordienko,** Departamento de Matemáticas, Universidad Autónoma Metropolitana-Iztapalapa. San Rafael Atlixco 186, Col. Vicentina, Iztapalapa, 09340, Mexico City, Mexico, e-mail: gord@xanum.uam.mx
Juan Ruiz de Chavez: Departamento de Matemáticas, Universidad Autónoma Metropolitana-Iztapalapa. San Rafael Atlixco 186, Col. Vicentina, Iztapalapa, 09340, Mexico City, Mexico, e-mail: jrch@xanum.uam.mx

In this article, we consider Markov control processes with general state and action spaces, choosing the expected total discounted reward as an optimization criterion. Thus, the results given in Section 3 are related to those obtained in the previous articles [9–11]. In contrast to the problem setting in these articles, we focus our attention on the controlled processes with *bounded* one-step rewards. This allows us to obtain new stability inequalities using both the total variation metric and the Dudley metric.

The total variation distance works well under the standard compactness-continuity conditions, but to obtain the corresponding stability inequality in terms of the Dudley metric, we have to impose additional Lipschitz continuity conditions.

The Dudley metric is convenient in an important situation where the nonparametric approach is applied, i.e., when unknown probability distributions are approximated by empirical distributions (see, e.g., [12]).

It should be noted that the problem of estimating the stability of optimal control considered in the article is closely related to the problem of adaptive control of Markov processes. In the adaptive formulation, the control is accompanied by some estimation procedure, and the current control policies should approximate the optimal ones as the distribution (or parameters) is refined. For the development of adaptive algorithms, quantitative estimates of the "stability of optimal control" can be useful. Among the vast literature, the works [13–17] used the expected total discounted reward as a criterion of optimization and discuss the application of nonparametric estimation of "governing distributions."

2 Setting of the problem

We consider a discrete-time Markov controlled process of the form:

$$X_t = F(X_{t-1}, a_t, \xi_t), \quad t = 1, 2, \dots, \tag{2.1}$$

where $X_t \in \mathfrak{X}$ is *a state* of the process at time t, and $\xi_1, \xi_2, \dots$ is a sequence of independent and identically distributed (i.i.d.) random vectors with values in a complete separable metric space $(\mathbf{S}, \rho)$. Let A be a *given action* set. Then, if $X_{t-1} = z \in \mathfrak{X}$, *then the control* (action) a_t is selected from a designated *compact* subset $A(z) \subset A$. We assume that $\mathfrak{X}$ and A are complete separable metric spaces (which are, particularly, Borel spaces). The metric in $\mathfrak{X}$ will be denoted by d. Finally, $F : \mathfrak{X} \times A \times \mathbf{S} \to \mathfrak{X}$ is a measurable function.

A sequence $\pi = (a_1, \dots, a_t, \dots)$, where the control a_t at time t is a measurable function of the current state x_{t-1} and can also depend on previous states and actions, is called *control policy*, or simply *policy*. A policy π is called *stationary* and denoted by f if there is a measurable function $f : \mathfrak{X} \to A$ such that $a_t = f(X_{t-1}) \in A(X_{t-1})$, $t = 1, 2 \dots$.

We denote by:

- Π the set of all policies;
- $\mathbb{F}$ the set of all stationary policies.

A policy optimization criterion, in our setting, is *the expected total discounted reward*:

$$V(x, \pi) = \mathbb{E}_x^{\pi} \sum_{t=1}^{\infty} \alpha^{t-1} r(X_{t-1}, a_t), \quad \pi \in \Pi, \quad x \in \mathfrak{X}, \tag{2.2}$$

where $\mathbb{E}_x^{\pi}$ is the expectation with respect to the probability that corresponds to the application of a policy π with an *initial state* of the process $x \in \mathfrak{X}$ (see, e.g., [18] for the construction of the corresponding probability space). $r(z, a)$ is *the one-step reward* acquired when the process is in the state z and the action a is selected and $\alpha \in (0, 1)$ is a given *discount factor*.

Throughout the article, we will assume that r is a measurable *bounded* function, that is,

$$\sup_{(x,a) \in \mathbb{K}} |r(x, a)| \le b < \infty. \tag{2.3}$$

In this inequality and further on, $\mathbb{K} \stackrel{\text{def}}{=} \{(x, a) \in \mathfrak{X} \times A : a \in A(x) \text{ for all } x \in \mathfrak{X}\}$, which is supposed to be a measurable subset of $\mathfrak{X} \times A$.

The policy π_* is called *optimal*, if for *each* $x \in \mathfrak{X}$,

$$\mathrm{V}(x, \pi_*) = \mathbb{V}_*(x) \stackrel{\text{def}}{=} \sup_{\pi \in \Pi} \mathrm{V}(x, \pi), \quad x \in \mathfrak{X}. \tag{2.4}$$

In many applications, all components of the process, *except* for the distribution G of the random vector ξ_1 in (2.1), are known. For the distribution G, usually some approximation $\tilde{G}$ is available (e.g., obtained from statistical data).

Despite the fact that a controller is looking for the optimal policy π_*, she/he is forced to work with the following *approximating controlled processes*:

$$\tilde{X}_t = F(\tilde{X}_{t-1}, \tilde{a}_t, \tilde{\xi}_t), \quad t = 1, 2, \ldots. \tag{2.5}$$

The only difference between this process and the "original" process in (2.1) is that the i.i.d. random vectors, $\tilde{\xi}_1, \tilde{\xi}_2, \ldots$, have the common distribution $\tilde{G}$.

The expected total discounted reward $\tilde{V}(x, \pi)$ for the process (2.5) is defined by formula (2.2), in which X_{t-1}, a_t is replaced by $\tilde{X}_{t-1}, \tilde{a}_t$.

Let $\mathbb{B}$ denote the space of all measurable *bounded* functions $u : \mathfrak{X} \to \mathbb{R}$, with the *uniform norm*:

$$\|u\| \stackrel{\text{def}}{=} \sup_{x \in \mathfrak{X}} |u(x)|.$$

Let ξ and $\tilde{\xi}$ be generic vectors for $\xi_1, \xi_2, \ldots$ and $\tilde{\xi}_1, \tilde{\xi}_2, \ldots$, respectively.

Assumption 1. For each fixed $x \in \mathfrak{X}$:
(a) the function $r(x, \cdot)$ is continuous on $A(x)$;
(b) for every $u \in \mathbb{B}$, the maps

$$a \to \mathbb{E}u[F(x, a, \xi)] \quad \text{and} \quad a \to \mathbb{E}u[F(x, a, \tilde{\xi})]$$

are continuous on $A(x)$.

The next assertion is well-known (see, e.g., [13, Ch. 2], and [19] for the proof).

Proposition 2.1. *Under Assumption* 1, *there exist stationary policies* $\pi_* \equiv f_*$ *and* $\tilde{\pi}_* \equiv \tilde{f}_*$, *which are optimal for the "real" process* (2.1) *and for the approximating process* (2.5), *respectively*.

In other words, (2.4) *holds with* f_*, *and also*

$$\tilde{\mathrm{V}}(x, \tilde{f}_*) = \tilde{\mathbb{V}}_*(x) \stackrel{\text{def}}{=} \sup_{\pi \in \Pi} \tilde{\mathrm{V}}(x, \pi), \quad x \in \mathfrak{X}. \tag{2.6}$$

Remark 2.1. If we assume that for a compact A, $A(x) = A$, $x \in \mathfrak{X}$ and the one-step reward function $r(x, a)$ is continuous on $\mathfrak{X} \times A$, then Proposition 2.1 holds true if we replace Assumption 1(b) with the following less restrictive condition:

Assumption 1. (b*): For each $x \in \mathfrak{X}$ and for every *continuous* and bounded function $u : \mathfrak{X} \to \mathbb{R}$, the maps

$$a \to \mathbb{E}u[F(x, a, \xi)] \quad \text{and} \quad a \to \mathbb{E}u[F(x, a, \tilde{\xi})]$$

are continuous on A. (See [19] for the corresponding proof of Proposition 2.1.)

Assume that the controller can find the policy $\tilde{f}_*$, and she/he applies $\tilde{f}_*$ to control the "original" process (2.1). In this way, $\tilde{f}_*$ is used as reasonable approximation to the not available policy f_*. We will measure the accuracy of such an approximation by evaluating the following *stability index*:

$$\Delta(x) \stackrel{\text{def}}{=} \mathrm{V}(x, f_*) - \mathrm{V}(x, \tilde{f}_*) \geq 0, \quad x \in \mathfrak{X}. \tag{2.7}$$

The problem under consideration is to prove stability inequalities of the type:

$$\sup_{x\in\mathfrak{X}}\Delta(x)\le C\boldsymbol{\mu}(G,\tilde{G}),$$

where $\boldsymbol{\mu}$ is either the total variation metric or the Dudley metric.

3 The results

First, we recall the definitions of two metrics in the space of distributions of random vectors with values in $(\mathbf{S},\mathcal{B}_{\mathbf{S}})$. Here, $\mathcal{B}_{\mathbf{S}}$ is the Borel σ-algebra of subsets of $\mathbf{S}$.

The total variation metric $\mathbb{V}$ (see, e.g., [20]):

If ξ and $\tilde{\xi}$ are random vectors with distributions G and $\tilde{G}$, then

$$\mathbb{V}(G,\tilde{G})\stackrel{\text{def}}{=}\sup_{\varphi\in\mathbb{B}_1}|\mathbb{E}\varphi(\xi)-\mathbb{E}\varphi(\tilde{\xi})|, \tag{3.1}$$

where

$$\mathbb{B}_1=\left\{\varphi:\mathbf{S}\to\mathbb{R}:\varphi\ \text{is measurable and}\ \|\varphi\|=\sup_{s\in\mathbf{S}}|\varphi(s)|\le 1\right\}.$$

The Dudley metric $\mathbf{d}$ (see [21]):

$$\mathbf{d}(G,\tilde{G})\stackrel{\text{def}}{=}\sup_{\varphi\in\mathbb{B}_{1,L}}|\mathbb{E}\varphi(\xi)-\mathbb{E}\varphi(\tilde{\xi})|, \tag{3.2}$$

where

$$\mathbb{B}_{1,L}=\left\{\varphi\in\mathbb{B}_1:\|\varphi\|+\sup_{s\ne s'}\frac{|\varphi(s)-\varphi(s')|}{\rho(s,s')}\le 1\right\}, \tag{3.3}$$

where ρ is the metric in $\mathbf{S}$.

It is well-known that the convergence in the metric $\mathbf{d}$ is equivalent to the weak convergence of distributions (see, e.g., [21]).

Theorem 1. *Under* (2.3) *and Assumption* 1,

$$\sup_{x\in\mathfrak{X}}\Delta(x)\le\frac{2ab}{(1-\alpha)^2}\mathbb{V}(G,\tilde{G}). \tag{3.4}$$

Proof. In view of Proposition 2.1, we can write (2.4) and (2.6) as follows ($x\in\mathfrak{X}$):

$$\mathrm{V}(x,f_*)=\mathrm{V}_*(x)=\sup_{f\in\mathbb{F}}\mathrm{V}(x,f), \tag{3.5}$$

$$\tilde{\mathrm{V}}(x,f_*)=\tilde{\mathrm{V}}_*(x)=\sup_{f\in\mathbb{F}}\tilde{\mathrm{V}}(x,f). \tag{3.6}$$

Then, for arbitrary $x\in\mathfrak{X}$ by (2.7) and (3.5) and (3.6),

$$\begin{aligned}\Delta(x)&\le|\mathrm{V}(x,f_*)-\tilde{\mathrm{V}}(x,\tilde{f}_*)|+|\tilde{\mathrm{V}}(x,\tilde{f}_*)-\mathrm{V}(x,\tilde{f}_*)|\\&=|\sup_{f\in\mathbb{F}}\mathrm{V}(x,f)-\sup_{f\in\mathbb{F}}\tilde{\mathrm{V}}(x,f)|+|\mathrm{V}(x,\tilde{f}_*)-\tilde{\mathrm{V}}(x,\tilde{f}_*)|\\&\le 2\sup_{f\in\mathbb{F}}|\mathrm{V}(x,f)-\tilde{\mathrm{V}}(x,f)|.\end{aligned} \tag{3.7}$$

Let us fix an arbitrary stationary policy $f \in \mathbb{F}$ and define two operators:

$T_f : \mathbb{B} \to \mathbb{B}$ and $\tilde{T}_f : \mathbb{B} \to \mathbb{B}$ as follows ($u \in \mathbb{B}$):

$$T_f u(x) \stackrel{\text{def}}{=} \{r(x, f(x)) + \alpha \mathbb{E} u[F(x, f(x), \xi)]\}, \quad x \in \mathfrak{X}, \tag{3.8}$$

$$\tilde{T}_f u(x) \stackrel{\text{def}}{=} \{r(x, f(x)) + \alpha \mathbb{E} u[F(x, f(x), \tilde{\xi})]\}, \quad x \in \mathfrak{X}. \tag{3.9}$$

The following two facts are well-known (see, e.g., [13, Ch. 2]):

(a) The functions $V_f(\cdot) \stackrel{\text{def}}{=} V(\cdot, f)$ and $\tilde{V}_f(\cdot) \stackrel{\text{def}}{=} \tilde{V}(\cdot, f)$ (where "$\cdot$" stands for $x \in \mathfrak{X}$) belong to $\mathbb{B}$, and moreover, they are fixed points of the operators T_f and $\tilde{T}_f$, that is,

$$T_f V_f = V_f \quad \text{and} \quad \tilde{T}_f \tilde{V}_f = \tilde{V}_f. \tag{3.10}$$

(b) The operators T_f and $\tilde{T}_f$ are contractive with modulus α, that is, ($u, v \in \mathbb{B}$):

$$\|T_f u - T_f v\| \le \alpha \|u - v\|; \quad \|\tilde{T}_f u - \tilde{T}_f v\| \le \alpha \|u - v\|. \tag{3.11}$$

Therefore,

$$\|V_f - \tilde{V}_f\| = \|T_f V_f - \tilde{T}_f \tilde{V}_f\| \le \|T_f V_f - T_f \tilde{V}_f\| + \|T_f \tilde{V}_f - \tilde{T}_f \tilde{V}_f\| \le \alpha \|V_f - \tilde{V}_f\| + \|T_f \tilde{V}_f - \tilde{T}_f \tilde{V}_f\|.$$

Hence,

$$\|V_f - \tilde{V}_f\| \le \frac{1}{1 - \alpha} \|T_f \tilde{V}_f - \tilde{T}_f \tilde{V}_f\|. \tag{3.12}$$

Let us estimate the second factor on the right side of (3.12). By (3.8) and (3.9), we have

$$\|T_f \tilde{V}_f - \tilde{T}_f \tilde{V}_f\| = \alpha \sup_{x \in \mathfrak{X}} |\mathbb{E} \tilde{V}_f[F(x, f(x), \xi)] - \mathbb{E} \tilde{V}_f[F(x, f(x), \tilde{\xi})]|. \tag{3.13}$$

Using the definition of $\tilde{V}_f$ (i.e., (2.2) with $\tilde{X}_t$, $\tilde{a}_t$), we see that

$$\sup_{x \in \mathfrak{X}} |\tilde{V}_f(x)| \le \sum_{t=1}^{\infty} \alpha^{t-1} b = \frac{b}{1 - \alpha}. \tag{3.14}$$

Thus, for each x fixed, in (3.13), the function $\tilde{V}_f[F(x, f(x), s)]$ of $s \in \mathbf{S}$ is bounded by $b(1 - \alpha)^{-1}$. Applying the definitions (3.1), (3.13), and (3.12), we find that

$$\sup_{x \in \mathfrak{X}} |V_f(x) - \tilde{V}_f(x)| \le \frac{\alpha b}{(1 - \alpha)^2} \mathbb{V}(G, \tilde{G}).$$

Combining the last inequality with (3.7), we obtain (3.4). □

In a fairly common situation, the unknown distribution G is estimated by the empirical distribution $\tilde{G}_n$, obtained from the sample $\xi_1, \xi_2, \ldots, \xi_n$. Excluding the cases of discrete G, $\mathbb{V}(G, \tilde{G}_n)$ fails to approach zero as $n \to \infty$. Thus, in many situations, inequality (3.4) is useless. On the other hand, under mild conditions, we have:

$$\mathbf{d}(G, \tilde{G}_n) \to 0 \text{ almost surely}, \quad \text{and} \quad \mathbb{E}\mathbf{d}(G, \tilde{G}_n) \to 0 \quad \text{as } n \to \infty,$$

(see the end of this section.)

To obtain the stability inequality with the Dudley metric $\mathbf{d}$ on the right-hand side, we need additional Lipschitz conditions.

Assumption 2.

(a) There exist a constant L_0 and a measurable function $\overline{L}_1 : \mathbf{S} \to [0, \infty)$ such that:

$$(1) \quad |r(x, a) - r(y, a)| \le L_0 d(x, y), \quad \text{for all } (x, a), (y, a) \in \mathbb{K}; \tag{3.15}$$

$$(2)\quad d[F(x,a,\xi),F(y,a,\xi)] \le \overline{L}_1(\xi)d(x,y), \quad \text{for all } (x,a),(y,a) \in \mathbb{K}, \tag{3.16}$$

$$\mathbb{E}\overline{L}_1(\xi) = L_1 \quad \text{and} \quad \alpha L_1 < 1.$$

(b) There is a constant $L < \infty$ such that for each $(x,a) \in \mathbb{K}$; $s, s' \in \mathbf{S}$,

$$d[F(x,a,s),F(x,a,s')] \le L\rho(s,s'). \tag{3.17}$$

(c) A is compact and $A(x) = A$ for all $x \in \mathfrak{X}$.

Theorem 2. *Under Assumptions* 1 *and* 2,

$$\sup_{x\in\mathfrak{X}} \Delta(x) \le \frac{2\alpha}{(1-\alpha)^2}\left[\frac{b}{1-\alpha} + \frac{L_0 L}{1-\alpha L_1}\right]\mathbf{d}(G,\tilde{G}), \tag{3.18}$$

where $\mathbf{d}$ *is the Dudley metric defined in* (3.2).

Proof. We define the operators $T : \mathbb{B} \to \mathbb{B}$ and $\tilde{T} : \mathbb{B} \to \mathbb{B}$ as follows ($u \in \mathbb{B}$):

$$Tu(x) \stackrel{\text{def}}{=} \sup_{a\in A}\{r(x,a) + \alpha\mathbb{E}u[F(x,a,\xi)]\}, x \in \mathfrak{X}, \tag{3.19}$$

$$\tilde{T}u(x) \stackrel{\text{def}}{=} \sup_{a\in A}\{r(x,a) + \alpha\mathbb{E}u[F(x,a,\tilde{\xi})]\}, x \in \mathfrak{X}. \tag{3.20}$$

In [13, Ch. 2], it was proved that:

(1) $$V_* = TV_* \quad \text{and} \quad \tilde{V}_* = \tilde{T}\tilde{V}_*, \tag{3.21}$$

where V_* and $\tilde{V}_*$ defined in (3.5) and (3.6) are value functions of the process (2.1) and of the process (2.5), respectively.

(2) Both operators T and $\tilde{T}$ are contractive (with respect to $\|\cdot\|$) with modulus α.

Let us define the number (generally belonging to $[0,\infty]$):

$$\mu(\xi,\tilde{\xi}) \stackrel{\text{def}}{=} \sup_{(x,a)\in\mathbb{K}} |\mathbb{E}V_*[F(x,a,\xi)] - \mathbb{E}V_*[F(x,a,\tilde{\xi})]|. \tag{3.22}$$

The first step in the proof is to establish the following inequality:

$$\sup_{x\in\mathfrak{X}} \Delta(x) \le \frac{2\alpha}{(1-\alpha)^2}\mu(\xi,\tilde{\xi}). \tag{3.23}$$

For $(x,a) \in \mathbb{K}$, let

$$H(x,a) \stackrel{\text{def}}{=} r(x,a) + \alpha\mathbb{E}V_*[F(x,a,\xi)], \tag{3.24}$$

$$\tilde{H}(x,a) \stackrel{\text{def}}{=} r(x,a) + \alpha\mathbb{E}\tilde{V}_*[F(x,a,\tilde{\xi})], \tag{3.25}$$

and for each $t \ge 1$,

$$\Gamma_t = \{x, a_1; X_1, a_2; \dots, X_{t-1}, a_t\}$$

be the part of a trajectory of the process (2.1) when applying the stationary policy $\tilde{f}_*$.

By Markov property of process (2.1) (when a stationary policy is applied) and (3.24), we have:

$$\zeta_t \stackrel{\text{def}}{=} \mathbb{E}^{\tilde{f}_*}[\alpha V_*(X_t)|\Gamma_t] = H(X_{t-1},a_t) - r(X_{t-1},a_t) = H(X_{t-1},a_t) - r(X_{t-1},a_t) - \sup_{a\in A} H(X_{t-1},a) + \sup_{a\in A} H(X_{t-1},a).$$

We can see from (3.24), (3.19), and (3.21) that

$$\sup_{a\in A} H(X_{t-1},a) = V_*(X_{t-1}).$$

Hence,

$$\zeta_t = H(X_{t-1}, a_t) - \sup_{a\in A} H(X_{t-1}, a) - r(X_{t-1}, a_t) + V_*(X_{t-1}) = -\Lambda_t - r(X_{t-1}, a_t) + V_*(X_{t-1}), \tag{3.26}$$

where

$$\Lambda_t \stackrel{\text{def}}{=} \sup_{a\in A} H(X_{t-1}, a) - H(X_{t-1}, a_t). \tag{3.27}$$

Now, rewriting (3.26) as

$$V_*(X_{t-1}) - r(X_{t-1}, a_t) - \zeta_t = \Lambda_t$$

and taking expectation $\mathbb{E}_x^{\tilde{f}_*}$ (in both parts), we obtain:

$$\mathbb{E}_x^{\tilde{f}_*} V_*(X_{t-1}) - \mathbb{E}_x^{\tilde{f}_*} r(X_{t-1}, a_t) - \alpha \mathbb{E}_x^{\tilde{f}_*} V_*(X_t) = \mathbb{E}_x^{\tilde{f}_*} \Lambda_t.$$

Multiplying the last equality by α^{t-1} and summing the inequalities with $t = 1, 2 \ldots, n$, we obtain:

$$V_*(x) - \alpha^n \mathbb{E}_x^{\tilde{f}_*} V_*(X_n) - \sum_{t=1}^{n} \alpha^{t-1} \mathbb{E}_x^{\tilde{f}_*} r(X_{t-1}, a_t) = \sum_{t=1}^{n} \alpha^{t-1} \mathbb{E}_x^{\tilde{f}_*} \Lambda_t. \tag{3.28}$$

From (3.14), it follows that V_* is a bounded function. So, taking in (3.28) limit $n \to \infty$, the second term on the left-hand side tends to zero, while the third term approaches $V(x, \tilde{f}_*)$. Therefore,

$$\Delta(x) = V_*(x) - V(x, \tilde{f}_*) = \sum_{t=1}^{\infty} \alpha^{t-1} \mathbb{E}_x^{\tilde{f}_*} \Lambda_t. \tag{3.29}$$

Since $\tilde{f}_*$ is the optimal policy for the process (2.5) applying (3.20), (3.21), and (3.25), we easily find that

$$\sup_{a\in A} \tilde{H}(X_{t-1}, a) = \tilde{H}(X_{t-1}, a_t).$$

Hence, by (3.27),

$$\Lambda_t = \sup_{a\in A} H(X_{t-1}, a) - \sup_{a\in A} \tilde{H}(X_{t-1}, a) + \tilde{H}(X_{t-1}, a_t) - H(X_{t-1}, a_t)$$

and

$$|\Lambda_t| \le \sup_{a\in A} 2|H(X_{t-1}, a) - \tilde{H}(X_{t-1}, a)| \le 2\alpha \sup_{a\in A} |\mathbb{E} V_*[F(X_{t-1}, a, \xi)] - \mathbb{E} \tilde{V}_*[F(X_{t-1}, a, \tilde{\xi})]|,$$

where the expectation in the last term is taken with respect to the random vectors ξ and $\tilde{\xi}$ (with X_{t-1} being fixed). From the last inequality, we obtain:

$$|\Lambda_t| \le 2\alpha \sup_{a\in A} |\mathbb{E} V_*[F(X_{t-1}, a, \xi)] - \mathbb{E} V_*[F(X_{t-1}, a, \tilde{\xi})]| + 2\alpha \sup_{a\in A} |\mathbb{E} V_*[F(X_{t-1}, a, \tilde{\xi})] - \mathbb{E} \tilde{V}_*[F(X_{t-1}, a, \tilde{\xi})]|. \tag{3.30}$$

The first term on the right-hand side of (3.30) is not greater than $2\alpha\mu(\xi, \tilde{\xi})$ (see (3.22)), and the second term is not greater than $2\alpha\|V_* - \tilde{V}_*\|$.

Using (3.21) and the contractive property of $\tilde{T}$, we have

$$\|V_* - \tilde{V}_*\| \le \|\tilde{T}\tilde{V}_* - \tilde{T}V_*\| + \|\tilde{T}V_* - TV_*\| \le \alpha\|V_* - \tilde{V}_*\| + \|TV_* - \tilde{T}V_*\|$$

or (see (3.19), (3.20))

$$\|V_* - \tilde{V}_*\| \le \frac{\alpha}{1-\alpha} \sup_{x\in\mathfrak{X}} \sup_{a\in A} |\mathbb{E} V_*[F(x, a, \xi)] - \mathbb{E} V_*[F(x, a, \tilde{\xi})]| \le \frac{\alpha}{1-\alpha} \mu(\xi, \tilde{\xi}).$$

The last inequality and (3.30) provide that for each $t \ge 1$,

$$|\Lambda_t| \le 2\alpha\left(1 + \frac{\alpha}{1-\alpha}\right)\mu(\xi, \tilde{\xi}).$$

Substituting this in (3.29), we obtain (3.23).

The second step in the proof of the theorem is to show that under Assumption 2, in (3.23),

$$\mu(\xi, \tilde{\xi}) \le \left[\frac{b}{1-\alpha} + \frac{L_0 L}{1-\alpha L_1}\right]\mathbf{d}(G, \tilde{G}). \tag{3.31}$$

By (3.14), the function V_* in (3.22) is bounded by $b(1-\alpha)^{-1}$. Now, we will show that for all $(x, a) \in \mathbb{K}$; $s, s' \in \mathbf{S}$,

$$|V_*[F(x, a, s)] - V_*[F(x, a, s')]| \le \tilde{L}\rho(s, s'), \tag{3.32}$$

$$\text{where } \tilde{L} = \frac{L_0 L}{1-\alpha L_1}. \tag{3.33}$$

First, we check that the value function $V_* : \mathfrak{X} \to \mathbb{R}$ satisfies the Lipschitz conditions with the constant $L_0/(1-\alpha L_1)$.

Let $u_0 \equiv 0$ and T be the operator defined in (3.19). Also, set $u_1 = Tu_0$. Then, for any $x, y \in \mathfrak{X}$,

$$|u_1(x) - u_1(y)| = |\sup_{a\in A} r(x, a) - \sup_{a\in A} r(y, a)| \le \sup_{a\in A}|r(x, a) - r(y, a)| \le L_0 d(x, y), \tag{3.34}$$

due to (3.15) in Assumption 2.

Let now $u_2 = Tu_1$. Then, in view of (3.19),

$$\begin{aligned}|u_2(x) - u_2(y)| &\le \sup_{a\in A}\{|r(x, a) - r(y, a)| + \alpha|\mathbb{E}u_1[F(x, a, \xi)] - \mathbb{E}u_1[F(y, a, \xi)]|\}\\ &\le L_0 d(x, y) + \alpha L_0 \sup_{a\in A}\mathbb{E}r(F(x, a, \xi), F(y, a, \xi))\\ &\le L_0(1 + \alpha L_1)d(x, y).\end{aligned}$$

To obtain the last inequality, we have made use of (3.34) and Assumption 2(a), (2).

Letting $u_n = Tu_{n-1}$, $n \ge 1$, it is proved by induction that for any x, y,

$$|u_n(x) - u_n(y)| \le L_0[1 + \alpha L_1 + \ldots,(\alpha L_1)^{n-1}]d(x, y). \tag{3.35}$$

Since V_* is a fixed point of the contractive operator T, we have $\|V_* - T^n u_0\| \to 0$ as $n \to \infty$. We see from (3.35) that for every $n \ge 1$, the function $u_n = T^n u_0$ is Lipschitz with the constant $\tilde{\tilde{L}} = \frac{L_0}{(1-\alpha L_1)}$. Consequently, V_* satisfies the Lipschitz condition with the constant $\tilde{\tilde{L}}$.

To verify (3.32), observe that by Assumption 2(b), the function $\varphi(s) = V_*[F(x, a, s)]$ is a composition of two Lipschitz functions.

Note that $\|\varphi\| \le b(1-\alpha)^{-1}$, and φ is Lipschitz with the constant $\tilde{L}$ in (3.33). Therefore, if we divide φ by $b(1-\alpha)^{-1} + \tilde{L}$, we obtain a function from the class $\mathbb{B}_{1,L}$ in (3.3). Finally, to obtain inequality (3.31), it suffices to compare (3.22) with the definition of the Dudley metric given in (3.2) and (3.3). □

The natural question arises: How to evaluate $\mathbf{d}(G, \tilde{G})$ in (3.18) if the distribution G is assumed to be unknown? We can give the answer in one of the most important cases when $\tilde{G}$ is the empirical distribution, used to estimate G.

Now, we assume that the random vectors $\xi_1, \xi_2 \ldots$ in (2.1) are observable and let $\xi_1, \xi_2 \ldots, \xi_n$ be i.i.d. observations of a random vector ξ with distribution G. The *empirical distribution* $\tilde{G} \equiv \tilde{G}_n$ is defined (on $(\mathbf{S}, \mathcal{B}_\mathbf{S})$) as follows:

$$\tilde{G}_n = \frac{1}{n}\sum_{k=1}^{n}\delta_{\xi_k}, \quad \text{where for } k = 1, 2 \ldots, n,$$

$$\delta_{\xi_k}(B) = \begin{cases}1, & \text{if } \xi_k \in B,\\ 0, & \text{otherwise}.\end{cases}$$

$(B \in \mathcal{B}_\mathbf{S})$.

Assume that $\mathbf{S} = \mathbb{R}^k$ and $|\cdot|$ is the Euclidian norm. Also, suppose that there exist constants $K < \infty$ and $h > 0$ such that $\mathbb{E}e^{h|\xi|} \le K$.

Then, there is a calculable constant $C = C(k, K, h)$ such that: for each $n = 1, 2, \ldots$

$$\mathbb{E}\mathbf{d}(G, \tilde{G}_n) \le C\delta(k, n), \tag{3.36}$$

$$\text{where } \delta(k, n) = \begin{cases} \dfrac{\log(1+n)}{n^{1/2}}, & \text{if } k = 1, \\ \dfrac{\log^2(1+n)}{n^{1/2}}, & \text{if } k = 2, \\ \dfrac{\log(1+n)}{n^{1/k}}, & \text{if } k \ge 3. \end{cases}$$

The inequality (3.36) was shown in Proposition 2.1 in [10], but it is actually a fairly direct consequence of Proposition 3.4 in [12]. Taking expectation in both parts of (3.18) one can apply inequality (3.36).

Remark 3.1. There is a class of controlled Markov processes with observable "perturbations" $\xi_1, \xi_2 \ldots$. (One representative is discussed in Example 2.) Even more often, the mentioned random vectors are not observable. In such cases, one should either use some indirect methods of bounding $\mathbf{d}(G, G_n)$, or look for other treatments. It is worth noting that our setting of the problem, generally speaking, does not require any estimation procedure. The distribution G can be, for example, some "theoretical simplification" of the known, but "too complex" real distribution G.

4 Examples

Example 1. In fact, this is a simple counterexample showing that Assumption 2 is essential for inequality (3.18) to hold.

Let $\mathfrak{X} = [0, \infty)$, $A = \{0, 1\}$, $\mathbf{S} = \mathbb{R}^2$, and for $\xi_t = (\xi_t^{(1)}, \xi_t^{(2)})$,

$$X_t = \xi_t^{(1)} + X_{t-1}a_t\xi_t^{(2)}, \quad t = 1, 2, \ldots.$$

For $a = 0$ and $a = 1$, the one-step reward function is the same and given by the following formula:

$$r(x, a) = \begin{cases} 2, & \text{if } x = 0 \\ x, & \text{if } x \in (0, 1] \\ 1, & \text{if } x > 1. \end{cases} \tag{4.1}$$

For an arbitrary but fixed $\varepsilon > 0$, we set $G = \delta_{(0,1)}$, that is,

$$P(\xi_t^{(1)} = 0) = 1 \quad \text{and} \quad P(\xi_t^{(2)} = 1) = 1,$$

and also, $\tilde{G} = \delta_{(\varepsilon,1)}$, that is

$$P(\xi_t^{(1)} = \varepsilon) = 1 \quad \text{and} \quad P(\xi_t^{(2)} = 1) = 1.$$

Then, the "real" process is

$$X_t = X_{t-1}a_t, \quad t \ge 1, \tag{4.2}$$

and the approximating one is

$$\tilde{X}_t = \varepsilon + \tilde{X}_{t-1}a_t, \quad t \ge 1. \tag{4.3}$$

Let $\alpha \in (0, 1)$ be any discount factor.

Let us fix the initial state $X_0 = \tilde{X}_0 = 1$. From (4.1) and (4.2), we see that the optimal stationary policy for the process (4.2) is $f_* = \{0, 0, \ldots\}$ (i.e., always select the action $f_*(x) = 0$). The corresponding reward is

$$V(1, f_*) = 1 + \sum_{t=2}^{\infty} \alpha^{t-1} \cdot 2 = \frac{1}{1-\alpha} + \frac{\alpha}{1-\alpha}. \tag{4.4}$$

Since the process (4.3) can never reach the state $x = 0$ and $r(x, a)$ is non-decreasing on $(0, \infty)$, the stationary optimal policy for the process (4.3) is $\tilde{f}_* = \{1, 1, \dots\}$. The application of $\tilde{f}_*$ to the process (4.2) gives

$$V(1, \tilde{f}_*) = \sum_{t=1}^{\infty} \alpha^{t-1} \cdot 1 = \frac{1}{1-\alpha}.$$

Comparing this with (4.4), we see that the stability index in (2.7) is

$$\Delta(1) = \frac{\alpha}{1-\alpha} > 0.$$

On the other hand, it is easy to show that $\mathbf{d}(G, G_\varepsilon) = \varepsilon \to 0$ (as $\varepsilon \to 0$).

Note that $\mathbb{V}(G, G_\varepsilon) = 2$ for all $\varepsilon > 0$.

Example 2. (See, e.g., [13, Ch. 1] or [22].) In this model, related to a dam operation, the stocks of water are specified by the equations

$$X_t = \min\{X_{t-1} - a_t + \xi_t, M\}, \quad t = 1, 2, \dots, \tag{4.5}$$

where $M < \infty$ is the capacity of a water reservoir, X_{t-1} is *the stock of water* at the beginning of tth period (say, day). The control a_t is the volume of water released during the tth period (e.g., for irrigation). Finally, ξ_t is a non-negative random variable representing the water inflow in the tth period. We assume that $\xi_1, \xi_2, \dots$ are i.i.d. random variables having density g.

As we see from (4.5), for this control process, $\mathfrak{X} = [0, M]$, $A(x) = [0, x]$, $x \in [0, M]$, and $\mathbf{S} = [0, \infty)$.

Choosing some *bounded* one-step reward function $r(x, a)$ (which in the simplest case is $(-1)\times$ the cost of a unit of water) and fixing a discount factor $\alpha \in (0, 1)$, we are faced with the problem of optimal water management, which is set as maximizing the expected long-term total discounted reward.

We assume that the density g (of the water inflow) is unknown, and it is approximated by some known density $\tilde{g}$ (obtained, for instance, from statistical estimations).

Also, we assume the following:

- For each $x \in [0, M]$, the one-step reward $r(x, a)$ is a continuous function of $a \in [0, x]$.
- Both densities g and $\tilde{g}$ are bounded and continuous on $(0, \infty)$.

In the verification of Assumption 1, (b) is a matter of simple calculations. Then, according to Proposition 2.1, there exist stationary optimal policies f_* and $\tilde{f}_*$, for the process (4.5) and, correspondingly, for the following approximating water release process:

$$\tilde{X}_t = \min\{\tilde{X}_{t-1} - \tilde{a}_t + \tilde{\xi}_t, M\}, \quad t = 1, 2, \dots,$$

where the i.i.d. random variables have density $\tilde{g}$. The application, for instance, of the policy f_* signifies that at tth period, the part $f_*(X_{t-1})$ of a current stock X_{t-1} is released.

Noting that all conditions of Theorem 1 are satisfied, and for distributions having densities,

$$\mathbb{V}(G, \tilde{G}) = \int_0^{\infty} |g(y) - \tilde{g}(y)| \mathrm{d}y,$$

by (3.4) we have

$$\sup_{x \in [0,M]} \Delta(x) \le \frac{2\alpha b}{(1-\alpha)^2} \int_0^{\infty} |g(y) - \tilde{g}(y)| \mathrm{d}y,$$

where $b = \sup_{(x,a) \in \mathbb{K}} |r(x, a)|$, and $\Delta(x)$ is the stability index defined in (2.7).

Example 3. (*Controlled "environmental stochastic process"*) The uncontrolled version of this discrete-time stochastic processes is defined by following recurrent equations (see, e.g., [23, Ch. 9]):

$$X_t = \alpha(\xi_t)X_{t-1} + \varphi(\xi_t), \quad t = 1, 2, \dots, \tag{4.6}$$

where $\xi_1, \xi_2 \dots$ are i.i.d. random vectors with values in the Euclidian space $\mathbb{R}^k$, and $X_t \in \mathbb{R}$ ($t = 0, 1, 2, \dots$).

Processes of type (4.6) are used in modeling some phenomena in environmental science.

We will consider a controlled variant of (4.6), that is, the process

$$X_t = \alpha(\xi_t)X_{t-1} + \varphi(a_t, \xi_t), \quad t = 1, 2, \dots, \tag{4.7}$$

where $a_t \in A$, and A is a given compact subset of the Euclidian space $\mathbb{R}^m$.

In this way, $A(x) = A$ for all $x \in \mathfrak{X} = \mathbb{R}$. In this example, the space $\mathbf{S}$ is $\mathbb{R}^k$.

Let $r(x, a)$ be a certain, bounded by b, one-step reward function, which is continuous on $\mathbb{R} \times A$, and, moreover, for some $L_0 < \infty$,

$$|r(x, a) - r(y, a)| \le L_0|x - y|, \tag{4.8}$$

for all $x, y \in \mathbb{R}$ and $a \in A$.

We assume that

$$\mathbb{E}\alpha(\xi_1) \le L_1 \quad \text{and} \quad \alpha L_1 < 1, \tag{4.9}$$

and, for some $L < \infty$,

$$|\varphi(a, s) - \varphi(a, s')| \le L|s - s'|, \tag{4.10}$$

for all $s, s' \in \mathbb{R}^k$, and $a \in A$, and also for each $s \in \mathbb{R}^k$, the map $a \to \varphi(a, s)$ is continuous. Using (4.7)–(4.10), it is easy to check the fulfillment of Assumption 2. Also, Assumption 1(a) and (b*) are fulfilled. Indeed, if $u : \mathbb{R} \to \mathbb{R}$ is continuous and bounded, then the map $a \to \mathbb{E}u[\alpha(\xi)x + \varphi(a, \xi)]$ is continuous by the dominated convergence theorem.

All of the above allow us to apply the stability inequality (3.18). Making use of the known relationship between the Dudley and Wasserstein metric, for the particular case where $k = 1$ (i.e., ξ_t is a random variable), the mentioned inequality can be written as follows:

$$\sup_{x\in\mathbb{R}} \Delta(x) \le \frac{2^{3/2}\alpha}{(1-\alpha)^2}\left[\frac{b}{1-\alpha} + \frac{L_0 L}{1-\alpha L_1}\right]\left(\int_{-\infty}^{\infty} |F_\xi(y) - F_{\tilde\xi}(y)|\mathrm{d}y\right)^{1/2},$$

where F_ξ and $F_{\tilde\xi}$ are the distribution functions of ξ and $\tilde\xi$, respectively, and $\tilde\xi$ is generic for i.i.d. random vectors $\tilde\xi_1, \tilde\xi_2, \dots$ involved in the approximating process

$$\tilde X_t = \alpha(\tilde\xi_t)\tilde X_{t-1} + \varphi(\tilde a_t, \tilde\xi_t), \quad t = 1, 2, \dots.$$

Acknowledgement: We thank the reviewers for their careful revision of the manuscript and for suggestions, which allow us to correct and improve the presentation of the article.

Author contributions: All authors read and approved the final manuscript.

Conflict of interest: The authors state no conflict of interest.

Data availability statement: No data, models, or code are generated or used during the study.

References

[1] S. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag, London, 1993.

[2] Ch. Andrieu, V. B. Tadić, and M. Vihola, *On the stability of some controlled Markov chains and its applications to stochastic approximation with Markovian dynamic*, Ann. Appl. Probab. **25** (2015), no. 1, 1–45, DOI: https://doi.org/10.1214/13-AAP953.

[3] Y. F. Atchadé and G. Fort, *Limit theorems for some adaptive MCMC algorithms with subgeometric kernels: Part II*, Bernoulli **18** (2012), no. 3, 975–1001, DOI: https://doi.org/10.3150/11-BEJ360.

[4] V. M. Zolotarev, *On the continuity of stochastic sequences generated by recurrent processes*, Theory Probab. Appl. **20** (1975), no. 4, 819–832, DOI: https://doi.org/10.1137/1120088.

[5] N. V. Kartashov, *Inequalities in stability and ergodicity theorems for Markov chains with a general phase space. II*, Teor. Veroyatn. Primen. **30** (1985), no. 3, 478–485, (in Russian).

[6] V. V. Kalasnikov and S. A. Anichkin, *Continuity of random sequences and approximation of Markov chains*, Adv. Appl. Probab. **13** (1981), no. 2, 402–414.

[7] N. M. Van Dijk, *Perturbation theory for unbounded Markov reward processes with applications to queuing*, Adv. Appl. Probab. **20** (1988), no. 1, 99–111, DOI: https://doi.org/10.2307/1427272.

[8] E. I. Gordienko, *Stability estimates for controlled Markov chains with a minorant. Stability problems of stochastic models*, J. Sov. Math. **40** (1988), 481–486, DOI: https://doi.org/10.1007/BF01083641.

[9] R. Montes-de-Oca, A. Sakhanenko, and F. Salem-Silva, *Estimates for perturbations of general discounted Markov control chains*, Appl. Math. **30** (2003), no. 3, 287–304, DOI: https://doi.org/10.4064/am30-3-4.

[10] E. Gordienko, E. Lemus-Rodriiiiguez, and R. Montes-de-Oca, *Discounted cost optimality problem: Stability with respect to weak metrics*, Math. Methods Oper. Res. **68** (2008), no. 1, 77–96, DOI: https://doi.org/10.1007/s00186-007-0171-z.

[11] E. I. Gordienko and F. S. Salem, *Robustness inequality for Markov control processes with unbounded costs*, Systems Control Lett. **33** (1998), no. 2, 125–130, DOI: https://doi.org/10.1016/S0167-6911(97)00077-7.

[12] R. M. Dudley, *The speed of mean Glivenko-Cantelli convergence*, Ann. Math. Statist. **40** (1969), 40–50, DOI: https://doi.org/10.1214/aoms/1177697802.

[13] O. Hernández-Lerma, *Adaptive Markov Control Processes, Applied Mathematical Sciences*, vol. 79, Springer-Verlag, New York, 1989.

[14] M. Schäl, *Estimation and control in discounted stochastic dynamic programming*, Stochastics **20** (1987), no. 1, 51–71, DOI: https://doi.org/10.1080/17442508708833435.

[15] R. Cavazos-Cadena, *Nonparametric adaptive control of discounted stochastic systems with compact state space*, J. Optim. Theory Appl. **65** (1990), no. 2, 191–207, DOI: https://doi.org/10.1007/BF01102341.

[16] O. Hernández-Lerma and S. I. Marcus, *Adaptive control of discounted Markov decision chains*, J. Optim. Theory Appl. **46** (1985), no. 3, 227–235, DOI: https://doi.org/10.1007/BF00938426.

[17] E. I. Gordienko and J. A. Minjárez-Sosa, *Adaptive control for discrete-time Markov processes with unbounded costs: Discounted criterion*, Kybernetika (Prague) **34** (1998), no. 2, 217–234.

[18] K. Hinderer, Foundations of non-stationary dynamic programming with discrete time parameter, *Lecture Notes in Operations Research and Mathematical Systems*, Vol. 33, Springer-Verlag, Berlin-New York, 1970.

[19] O. Hernández-Lerma and M. Muñoz de Özak, *Discrete-time Markov control processes with discounted unbounded costs: optimality criteria*, Kybernetika (Prague) **28** (1992), no. 3, 191–212.

[20] S. T. Rachev, *Probability Metrics and the Stability of Stochastic Models*, John Wiley & Sons, Ltd., Chichester, 1991.

[21] R. M. Dudley, *Real analysis and probability*, in: Cambridge Studies in Advanced Mathematics, vol. 74, Cambridge University Press, Cambridge, 2002, Revised reprint of the 1989 original.

[22] S. Yakowitz, *Dynamic programming applications in water resources*, Water Resources **18** (1982), 673–696.

[23] S. T. Rachev and L. Rüschendorf, *Mass Transportation Problems. Vol. II: Applications*, Probability and its Applications (New York), Springer-Verlag, New York, 1998.