Stability estimation of Markov control processes with unbounded rewards

E. I. Gordienko [1], ***J. Ruiz de Chavez*** [2]

[1]Universidad Autonoma Metropolitana-I, Mexico City, Mexico, gord@xanum.uam.mx
[2]Universidad Autonoma Metropolitana-I, Mexico City, Mexico, jrch@xanum.uam.mx

The problem of stability estimation of Markov control processes considered here was, probably, for the first time set and solved for certain particular processes, in the papers Van Dijk and Puterman [1] and Gordienko [2]. It turned out that the method of probability metrics, for the most part developed in the works of V. M. Zolotarev (see e.g. Zolotarev [3]) has been very helpful to tackle problems of this type.

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

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

where:

- X_t is a *a state* of the process belonging to a separable metric space $\mathfrak{X}$;
- $\xi_1, \xi_2, \dots$ is a sequence of i.i.d. random vectors with values in a separable metric space S;
- when $X_{t-1} = x \in \mathfrak{X}$, $a_t \in A(x)$ is the *control* (action) at time t, that is selected from a given compact subset $A(x) \subset A$; here the action set A is also a separable metric space;
- finally, $F : \mathfrak{X} \times A \times S \to \mathfrak{X}$ is a given measurable function.

A sequence $\pi = (a_1, \dots, a_t, \dots)$, where the control a_t can depend on previous states and actions, is called *control policy* , or simply *policy*. We denote by Π the set of all policies. A policy $\pi \equiv f$ is called *stationary* if there is a measurable function $f : \mathfrak{X} \to A$ such that for each $t = 1, 2, \dots$, $a_t = f(X_{t-1}) \in A(X_{t-1})$.

The *optimal policy* π_* is such that provides a maximum value of a performance criteria, which in this talk is specified to be *an expected total discounted reward*:

$$V(x, \pi) = E_x^\pi \sum_{t=1}^{\infty} \alpha^{t-1} r(X_{t-1}, a_t), \tag{2}$$

where $\alpha \in (0, 1)$ is a given discount factor, and $r(x, a)$ is the *one-step reward* acquired when the process is in the state x and the action a is selected. We

allow the function $f : \mathfrak{X} \times A \to \mathbb{R}$ to be *unbounded*. In (2) $x \in \mathfrak{X}$ is the *initial state* of the process.

Let G be the distribution of the random vector ξ_t. In many applied controlled processes all components of the above model, excepting G, can be known. For the latter, commonly, some approximation $\tilde{G}$ (to G) is available. This approximation can be obtained by some theoretical speculations and /or statistical procedures.

In this way, the "real" control process (1) is unavailable for a researcher, and she/he should deal with its approximation:

$$\tilde{X}_t = F(\tilde{X}_{t-1}, \tilde{a}_t, \tilde{\xi}_t),\ t = 1, 2, \ldots, \tag{3}$$

where $\tilde{\xi}_1, \tilde{\xi}_2, \ldots$ are i.i.d. random vectors distributed according to $\tilde{G}$.

We let certain conditions (see, e.g. Hernandez-Lerma and Lasserre [4]) which ensure the existence of *optimal stationary policies* f_* and $\tilde{f}_*$ for the processes (1) and (3), respectively; that is:

$$V(x, f_*) = \sup_{\pi \in \Pi} V(x, \pi),\ x \in \mathfrak{X};$$

$$\tilde{V}(x, \tilde{f}_*) = \sup_{\pi \in \Pi} \tilde{V}(x, \pi),\ x \in \mathfrak{X},$$

where, similarly to (2),

$$\tilde{V}(x, \pi) = E_x^{\pi} \sum_{t=1}^{\infty} \alpha^{t-1} r(\tilde{X}_{t-1}, \tilde{a}_t).$$

The natural question arises: Is $\tilde{f}_*$ a "good approximation" to the not attainable optimal policy f_*? Having in the mind that the policy $\tilde{f}_*$ is intended in order to control the "original" process (1), we use the following *stability index*

$$\Delta(x) := V(x, f_*) - V(x, \tilde{f}_*) \geq 0,\ x \in \mathfrak{X}$$

as a measure of the quality of the approximation.

The problem of stability estimation settled here is establishing inequalities of the type:

$$\Delta(x) \leq B(x) \mu(G, \tilde{G}),\ x \in \mathfrak{X}, \tag{4}$$

where μ is a suitable probability metric. Taking into account a possible unboundedness of the one-step reward function $r(x, a)$, an appropriate candidate for μ in (4) is the, so-called, weighted total variation metric. In this talk supposing that $S = \mathbb{R}^k$ and assuming relevant moment conditions on G and $\tilde{G}$,

we present a version of *the stability inequality* (4) with $\mu = \mathbb{V}$ being the usual *total variation metric*.

Note, that under additional, rather restrictive Lipschitz conditions on the processes $\{X_t\}$ and $\{\tilde{X}_t\}$ a variant of (4), where μ is the Kantorovich metric, was proven in Gordienko et al [5].

References

1. N. M. Van Dijk, M. L. Puterman, Perturbation theory for Markov reward processes with applications to queueing systems, *Advances in Applied Probability* **20** , (1988) 79-98.

2. E. I. Gordienko, Stability estimates for controlled Markov chains with a minorant, *Journal Soviet Mathematics* **40:2** (1988) 481-486.

3. V. M. Zolotarev, Probability metrics, *Theory of Probability and its Applications* **28**:2 (1983) 264–287.

4. O. Hernández-Lerma, J. B. Lasserre, *Discrete-time Markov Control Processes*, Springer-Verlag, New York, 1995.

5. E. Gordienko, E. Lemus-Rodriguez, and R. Montes-de-Oca, Discounted cost optimality problem: stability with respect to weak metrics, *Mathematical Methods of Operations Research* **68:1** (2008) 77-96.