On Optimality in Probability and Almost Surely for Controlled Stochastic Processes with a Communication Property

Tatiana A. Belkina * and Vladimir Rotar

The Central Economics and Mathematics Institute of Russian Academy of Sciences (both authors);
Department of Mathematics and Statistics of San Diego State University, San Diego, CA 92182-7720, USA
(the second author)
E-mail: rotar@sciences.sdsu.edu, tbel@cemi.rssi.ru

*Supported in part by Russian Foundation of Basic Research, grant # 00-01-00194;
INITAS, 97-1050; The 6th Competition of Projects of Young Scholars (Russia), # 105.

January 9, 2002

Abstract

The paper concerns both controlled diffusion processes, and processes in discrete time. We establish conditions under which the strategy minimizing the expected value of a cost functional has a much stronger property; namely, it minimizes the random cost functional itself for all realizations of the controlled process from a set, the probability of which is close to one for large time horizons. The main difference of the conditions mentioned from those obtained earlier is that the former do not deal with strategies optimal in the mean themselves but concern a possibility of transition of the controlled process from one state to another in a time with a finite expectation. It makes the verification of these conditions in a number of situations much easier.

AMS 1991 Subject Classification: Primary 93E20, Secondary 49K45.

Keywords: controlled processes, optimal control, asymptotic optimality, optimality almost surely.
1 Introduction and some corollaries from the main theorems.

1.1 Background and motivation

Much of optimal stochastic dynamic control theory is concerned with minimization of the expectation of a cost functional over a class of admissible strategies. If the horizon is infinite, and the cost functional is additive, the notion of optimality is traditionally connected with the expected value of the long run average cost. In the sequel, we will call strategies of both types mentioned strategies \textit{optimal in the mean}.

In recent years, there has been a significant interest in the relatively new direction connected with an essentially stronger notion of optimality. This direction concerns the search of strategies which minimize not merely the expectation of the cost functional but the functional itself with a probability which is close to one asymptotically, for large time horizons. As a rule, if such strategies exist, among them there is also an optimal in the mean strategy.

To our knowledge, the first paper in this direction was that by Lippman (1971) where recurrent semi-Markov processes were considered. An essential advance was achieved in Mandl (1974a), Mandl (1974b), Mandl (1986), Mandl and Ayllon (1987); see therein references to other papers of the same authors. These papers concerned Markov controlled chains, and at the first time made use of a martingale approach for solving the problems under discussion. Leizarovitz (1988), Borkar (1989), Carlson, Haurie, and Leizarowitz (1991) dealt with homogeneous diffusion processes. In particular, for this case Borkar (1989) introduced into consideration the notion of a stable stationary control. The general discrete-time scheme with appropriate definitions of \textit{optimality almost surely} and \textit{in probability} was considered in Rotar (1986b), and Asriev and Rotar (1990). A general framework for diffusion processes was suggested in Presman, Rotar, Taksar (1993).

Since general conditions in all the papers mentioned are complicated, some particular schemes have been explored in more detail. Most, it is the linear model with quadratic costs; see, e.g., Chen and Guo (1987) for ARMAX model, Mandl (1974b), Hall and Heyde (1980, Chapter 7), Leizarovitz (1987), Konyukhova (Belkina) and Rotar (1992), Konyukhova (Belkina) (1994) for the linear regulator in discrete time, and Presman, Rotar, Taksar (1993) for the linear regulator in continuous time. For more comments see also Section 1.3.

It is worth noting that definitions of asymptotic optimality are somewhat different (and non-equivalent) in various papers; see Section 1.2.1 for some details. We follow here definitions of optimality in probability and almost surely introduced in Rotar (1986b), and Asriev and Rotar (1990).

General conditions in the papers mentioned above concern properties of the value (Bellman) function for minimizing the expected value of the cost functional. These conditions require that the value function does not strongly depend on the initial state of the process, as well as on the decisions made in early stages of control. More precisely, these factors should influence only the second order term in the asymptotic representation of the value function for large time horizons. It means, in particular, that in order to verify these conditions one should first find the strategy optimal in the mean, and - more or less explicitly - the value function itself. This is far from being always possible.

In this paper we consider assumptions of a different type which concern a \textit{communication
Consider a trajectory starting from a fixed initial state, and generated by a strategy optimal in the mean. We choose another initial state, and consider strategies, if any, such that the trajectories generated by these strategies, starting from the latter initial state, intersect the former trajectory at some (random) time. The condition under discussion assumes that among these strategies there exists a strategy for which the expected value of the coupling time mentioned is bounded uniformly with respect to initial states.

The main point here is that in order to verify such a condition, one does not need to know an explicit representation of the optimal in the mean trajectory. For instance, it suffices to verify a possibility of a transition from one state to another for a not “too long” time. The controlled finite Markov chain may serve as the simplest example. In this case, the above communication property holds if there exists an admissible strategy (perhaps, very “bad”) under which the process moves from any state to another state with, say, equal probabilities. Another simple illustration is the linear system with an arbitrary bounded cost per unit time: in this case the verification of the condition mentioned is also easy. The same is true for a much more general situation.

Thus, when the condition under discussion is verified, but it is difficult to find the optimal in the mean strategy explicitly, we can do that numerically, and can be still sure that this strategy is optimal not only in the mean but asymptotically almost surely.

For discrete time some (less general than in this paper) results in this direction were obtained in Rotar (1991).

In the rest of this section we consider some conditions and corollaries from the main theorems. The general situation is concerned in Section 2. Proofs are in Section 3.

1.2 Optimality almost surely for homogeneous processes.

1.2.1 Markov chains in discrete time.

A definition of optimality. We specify a homogeneous controlled Markov chain \( x_t, t = 0, 1, 2, \ldots \), with values in \( \mathbb{R}^l \), by the recurrence relation

\[
x_t = h(\xi_t, x_{t-1}, a_t),
\]

where \( \xi_1, \xi_2, \ldots \) are independent, identically distributed (i.i.d.) random variables (r.v.’s), \( a_t \) is an element from a space \( \mathbb{R}^m \), and a function \( h : \mathbb{R} \times \mathbb{R}^l \times \mathbb{R}^m \to \mathbb{R}^l \). We interpret \( a_t \) as an action at time \( t \).

(It is more traditional to specify a controlled chain by a transition measure \( Q(\cdot | x, a) \), the conditional distribution of \( x_t \), given \( x_{t-1} = x, a_t = a \). On the other hand, for each \( Q \) there exists a function \( h \) such that the distribution of the r.v. \( h(\xi_t, x, a) \) coincides with \( Q(\cdot | x, a) \). Moreover, the r.v.’s \( \xi \)’s may be assumed to be uniform on \( [0, 1] \); see, e.g., Gikhman and Skorokhod (1979). The representation (1.2.1) is more convenient for our purposes since the existence of a unique probability space allows to compare processes generated by different strategies.)

Below, for any sequence of elements \( \alpha_1, \alpha_2, \ldots \) we write \( \alpha' \) for \( (\alpha_1, \ldots, \alpha_t) \).
Set $x_0 = x$. For any integer $n$, consider a functional

$$J_{xn}(u^n) = \sum_{t=1}^{n} q(x_t, a_t),$$  

(1.2.2)

where $x_t$ satisfies (1.2.1) with $x_0 = x$, and a function $q$ specifies the cost per unit time.

To determine a control, one sets $a_t = u_t$, a r.v. measurable with respect to the $\sigma$-algebra $\mathcal{F}_{t-1} = \sigma\{\xi_1, \xi_2, \ldots, \xi_{t-1}\}$. The r.v. $u_t$ is said to be a policy (control) at time $t$.

The symbol $u$ will stand for an infinite $\{u_1, u_2, \ldots\}$, the symbol $u^n$ - for its restriction $\{u_1, \ldots, u_n\}$. The term policy applies to sequences $u$ too. The writing $J_{xn}(u)$ means $J_{xn}(u^n)$.

We fix a class of admissible policies $\mathcal{U}$, which is a class of infinite sequences mentioned.

Denote by $\{0, n\}$ the period $\{t; t = 0, \ldots, n\}$. A policy $\hat{u}^n$ is said to be optimal in the mean for the period $\{0, n\}$ and the initial state $x$ if

$$E \{ J_{xn}(\hat{u}^n) \} = \inf_{u^n} E \{ J_{xn}(u^n) \};$$

(1.2.3)

where inf is over all restrictions $u^n$ of policies from $\mathcal{U}$.

Clearly, $\hat{u}^n$ may depend on $x$ and $n$, so when considering a sequence $\{\hat{u}^n\}$ we deal with the triangular array $\{\hat{u}_{1n}, \ldots, \hat{u}_{nn}\}$. Later we will consider the general case when the control is specified in general by triangular arrays $\{u^n = (u_{1n}, \ldots, u_{nn})\}$; in this section we restrict ourselves to the traditional framework which includes only infinite sequences of policies $\{u_1, u_2, \ldots\}$.

**Definition 1** Given an initial state $x$, a policy $u^* \in \mathcal{U}$ is said to be asymptotically optimal almost surely if for any policy $u \in \mathcal{U}$

$$\liminf_{n \to \infty} n^{-1} (J_{xn}(u) - J_{xn}(u^*)) \geq 0 \quad \text{a.s.}$$  

(1.2.3)

The above definition does not exhaust all known definitions of such a type. One of well known schemes (see, e.g., Mandl (1974), Hall and Heyde (1980) for the linear regulator, Borkar (1989) for a general stationary model, Hernández-Lerma and Lasserre (1999) for some models in discrete time, and references in the papers mentioned) concerns conditions under which (a) there exists a policy $\overline{u}$ such that

$$\lim_{n \to \infty} n^{-1} J_{xn}(\overline{u}) = \theta \quad \text{a.s.};$$  

(1.2.4)

(b) for any admissible policy $u$

$$\limsup_{n \to \infty} n^{-1} J_{xn}(u) \geq \theta \quad \text{a.s.}$$  

(1.2.5)

The optimality property in the latter sense is weaker in general than that in (1.2.3) for two reasons. First, (1.2.3) does not presuppose a law of large numbers (LLN) as in (1.2.4), which in the general case - say, for a non-homogeneous scheme - may not take a place: the limit in the l.-h.s. may not exist even for “good” policies, or may equal not a number but a random variable. Speaking ahead, note that the last argument becomes essential if we consider a more general scheme than Markovian (which will be done in Section 2.2): in this case even for a homogeneous scheme the
limit in the LLN may occur random for “good” strategies and rather natural dependency structures [say, for exchangeable r.v.’s; see, e.g., Chow and Teicher (1997), Rotar (1997)].

Second, definitions (1.2.3) and (1.2.5) remains different when condition (1.2.4), or similar, does hold. Assume that for a policy $u^*$ the quantity $\liminf_{n \to \infty} n^{-1} J_{st}(u^*)$, certain or random, exists. Then (1.2.3) implies that

$$\liminf_{n \to \infty} n^{-1} J_{st}(u) \geq \liminf_{n \to \infty} n^{-1} J_{st}(u^*),$$

which is stronger, and more desirable if it is possible, than (1.2.5) since we deal now not with limsup but with liminf.

Below we follow Definition 1 because, as will be shown, policies with the property (1.2.3) exist under rather mild conditions.

The first proposition. The notion of a strategy is not necessary for the following but allows to simplify formulations. A strategy is a (finite or infinite) sequence $\psi = (\psi_1, \psi_2, \cdots)$, where $\psi_t : \mathbb{R}^{I(t-1)} \to \mathbb{R}^m$ specifies the policy at time $t$; namely, we set $u_t = \psi_t(x^{t-1})$ for $t \geq 1$. If $\psi_t(x^{t-1}) = \phi_t(x_{n-1})$ for any $t$, where $\phi_t : \mathbb{R}^I \to \mathbb{R}^m$, then the strategy is Markovian.

Denote by $U_{\psi_{st}}$, $X_{\psi_{st}}$, the values of the policy and the process, respectively, at time $t$, generated by a strategy $\psi$ and an initial state $x$. More precisely,

$$X_{\psi x 0} = x, \quad X_{\psi x t} = h(\xi_t, X_{\psi x (t-1)}, U_{\psi_{st}}) \quad \text{for} \quad t = 1, 2, \ldots,$$

$$U_{\psi x 1} = \psi_1(x), \quad U_{\psi_{st}} = \psi_t(X_{\psi x t-1}) \quad \text{for} \quad t = 1, 2, \ldots, \quad \text{and} \quad X_{\psi x t} = (X_{\psi x 0}, \cdots, X_{\psi x t}).$$

Set $U_{\psi x}^n = (U_{\psi x 1}, \cdots, U_{\psi x n})$. A strategy $\psi$ is said to be admissible if $(U_{\psi x 1}, U_{\psi x 2}, \cdots) \in \mathcal{U}$ for all $x$. The process generated by a strategy will be called a trajectory. When dealing with a trajectory, we will omit sometimes the index $\psi$ writing $U_{xt}$, $X_{xt}$.

If a strategy is defined only up to a time moment $n$, and the elements of this strategy perhaps depend on $n$, in order to emphasize it, we sometimes write $\psi_{tn}$, $nX_{xt}$, $nU_{xt}$. Such a strategy is called admissible if $U_{\psi x}^n$ is the restriction of a sequence from $\mathcal{U}$.

For a fixed $n$, a strategy $\Phi^n$ is said to be optimal in the mean if for each initial state $x$ the policy $nU_{\Phi x}$, which this strategy generates, is optimal in the mean for the period $\{0, n\}$. For brevity we denote such a policy by $n\hat{U}_x = (n\hat{U}_x, \cdots, n\hat{U}_x)$, and the corresponding trajectory - by $n\hat{X}_x = (n\hat{X}_x, \cdots, n\hat{X}_x)$.

Below we assume that for each $n$ a strategy optimal in the mean exists.

Proposition 2 Suppose $q(\cdot, \cdot)$ is bounded, and there exists a constant $C$ with the following property:

For any $n$ and an arbitrary pair of (initial) states $z$ and $z'$ there exist an admissible strategy $\phi = \Phi^n$ and a r.v. $N$ (perhaps depending on $n$, $z$ and $z'$) such that

$$E\{N\} \leq C, \quad \text{(1.2.7)}$$

and

$$N \leq t \leq n \Rightarrow q(X_{\phi z t}, U_{\phi z t}) = q(n\hat{X}_{z t}, n\hat{U}_{z t}). \quad \text{(1.2.8)}$$
Suppose also that for an initial state $x$ there exists a policy $\tilde{U}_x = \{\tilde{U}_x^n\}_{n=1}^{\infty} \in \mathcal{U}$ and such that
\[
\frac{1}{n} \left( J_{xn}(\tilde{U}_x^n) - J_{xn}(\bar{n}\tilde{U}_x^n) \right) \to 0 \quad \text{a.s., as} \quad n \to \infty. \tag{1.2.9}
\]

Then the policy $\tilde{U}_x$ is asymptotically optimal almost surely for the $x$ mentioned.

Next we discuss conditions of Proposition 2. Clearly, we need (1.2.9) merely because $\{\bar{n}\tilde{U}_x^n\}$ is a triangular array, that is, it is not from $\mathcal{U}$: otherwise we would call $\bar{n}\tilde{U}_x^n$ itself optimal a.s. Though in the most general case it is probably hard to establish conditions under which a policy $\tilde{U}_x$ with the property (1.2.9) exists, results for particular schemes show that these conditions can be rather mild. (See, e.g., Leizarovitz (1987), Borkar (1989), Presman, Rotar, Taksar (1993); Presman (1997), Di Masi, Kabanov (1998), and many other papers where the policy with the property mentioned is that minimizing the expected value of the long run average cost on $[0, \infty)$.)

Turn to the main conditions (1.2.7)-(1.2.8). Formally, for only (1.2.8) to hold, one can set, say, $N = n + 1$, since in this case the set of $t$’s in the l.-h.s. of (1.2.8) is empty and (1.2.8) holds automatically. However, such $N$ will not satisfy (1.2.7), so $N$ should not depend on $n$ “too much”. We can view $N$ as the time during which the process can reach - starting from an arbitrary fixed initial state $z$, and under a strategy we have a right to choose - the optimal in the mean trajectory starting from another initial state $z’$. We require the mean value of the coupling time to be bounded uniformly in $z, z’, n$.

The essential point here is the uniformity in $z$ and $z’, n$ being merely since the trajectory $\bar{n}\tilde{X}_z^n$ may depend on $n$. On the other hand, if there exists a policy $\bar{U}_x$ satisfying (1.2.9), then roughly speaking it suffices to think just about the time of a possible transition to the trajectory $\bar{n}\tilde{X}_z^n$.

It is worthwhile to note also that the strategy $\varphi^n$ is allowed to be, so to speak, “bad”, and condition (1.2.8) is a condition rather not on the process itself but on the class $\mathcal{U}$ of admissible strategies: it must contain the transition strategy mentioned.

For illustration, we briefly recall a very simple example from Asriev and Rotar (1990); for details see this paper. Let $x_t$ take values $-1, 0, 1$. Let states $\pm 1$ be absorbing under any strategy, and assume that if the process is at the state 0, two controls are possible: under the first the process remains at 0 with probability 1, while under the second control the process moves to $-1$ or 1 with probabilities $1/2$. Let the initial state $x = 0$, and $q(x_t, a_t) = x_t$. As is easy to see, in this case any policy is optimal in the mean, but there is no policy optimal almost surely.

The same is true in the general case: for a policy optimal a.s. to exist, the states should communicate in a sense. Certainly, in general we may require two trajectories from above to couple not exactly but only to approach each other (see Section 2), but some communication property must take place. In view of this, conditions for a.s. optimality suggested in this paper are close, in a certain sense, to minimal.

Next we consider two particular classical problems, rather not as examples of applications but as illustration of how the above conditions may be verified, and why the verification can be easy and does not require the search of particular policies, say, those optimal in the mean.

**EXAMPLE 1.** Denote by $p_m(\psi; n; z, z’)$ the probability that in $m$ steps the following two trajectories will intersect: the trajectory emanating from the state $z$ and generated by a strategy optimal in
At least for a finite chain condition (1.2.10) is rather weak. Assume, for instance, that the class of admissible strategies contains a Markov strategy \( \tilde{\varphi} \) under which the process moves from any state to another state with a probability greater than a fixed positive number. Say, under this strategy the system moves from any state to others with equal probabilities. Then (1.2.10) is true for \( \varphi = \tilde{\varphi} \), and \( m = 1 \).

Clearly, the condition of Proposition 2 would hold in the case (1.2.10). Indeed, for \( \varphi = (\varphi_1, \varphi_2, \ldots) \) from (1.2.8), we set \( \varphi_t = \varphi_t \) up to the moment of the coupling of the trajectories mentioned, and after this moment we control the process in accordance with the strategy optimal in the mean. Let \( N \) be the coupling time. For \( m \leq k < 2m \) we have \( P(N > k) \leq P(N > m) \leq 1 - q \). For \( lm \leq k < (l + 1)m \), we have \( P(N > k) \leq P(N > lm) = P(N > lm|N > (l - 1)m)P(N > (l - 1)m) \leq (1 - q)^2P(N > (l - 1)m) \). By induction, \( P(N > k) \leq (1 - q)^{k/m} \), which implies \( E\{N \} \leq m/q(1 - q) \). ■

EXAMPLE 2. Consider a controlled linear system

\[
x_t = Ax_{t-1} + Bu_t + \xi_t,
\]

where now \( \xi \)'s are i.i.d. \( l \)-dimensional random vectors, \( B \) is a matrix, policies \( u \)'s are random vectors, and the dimensions of \( B \) and \( u \) are such that the last relation is meaningful.

*The cost function \( g \) is still an arbitrary bounded function.*

We fix \( n \), and set

\[
y_t = n\tilde{X}_{t+1} - X_{t+1}, \quad V_t = n\tilde{U}_{t+1} - U_{t+1}, \quad t = 1, \ldots, n,
\]

(1.2.11)

where \( \varphi \) is a strategy. Clearly,

\[
y_t = Ay_{t-1} + BV_t,
\]

(1.2.12)

with \( y_0 = z' - z \).

It is known that, if the pair \((A, B)\) is the so called controllable pair, then for any initial state there exists a (non-random) policy \( V \) which transfers the system (1.2.12) from this state to the null state in \( l \) steps (see, e.g., Kwakernaak and Sivan (1972) for details). So, there exists \( V \) which transfers \( y_t \) from the state \( y_0 = z' - z \) to the null state.

(More precisely, one should set \( V_j = -B'(A')^{l-j-1}W_l^{-1}A^l(z' - z), \quad 0 \leq j \leq l - 1, \) and \( V_j = 0 \) for \( l \leq i \leq n \), where \( W_l = PP' \), and \( P = (B, AB, A^2B, \ldots, A^{l-1}B) \) is the so called matrix of controllability; see, e.g., Kwakernaak and Sivan (1972).)

With \( V \) chosen in this way, we choose \( \varphi \) in accordance with (1.2.11), and in this case (1.2.8) holds with \( N = l \). ■

Clearly, both above examples may be generalized in an essential way, though it could lead to more complicated calculations. We believe that Examples 1-2 are enough to illustrate the essence of the matter.
1.2.2 Continuous time: a homogeneous diffusion processes.

Consider a process \( x(t) \) with values in \( \mathbb{R}^l \), which is governed by the equation

\[
dx(t) = b(x(t), u(t))dt + \sigma(x(t), u(t))dw(t).
\]  

(1.2.13)

Here \( w(\cdot) \) is a standard \( d \)-dimensional Wiener process defined in a complete probability space \( (\Omega, \mathcal{F}, P) \); the symbol \( u(t) \) stands for an admissible policy, that is, \( u(t) \) is a random process with values in a space \( \mathbb{R}^m \), adapted to the family \( \mathcal{F}_t = \sigma(w(s), 0 \leq s \leq t) \); and \( b \) and \( \sigma \) are vector- and matrix-functions, respectively, of the corresponding dimensions. Set \( x(0) = x \).

Throughout the paper we assume that for any policy under consideration there exists a strong solution to (1.2.13), not specifying conditions under which it is true. (See also Section 2.1. On strong solutions and corresponding conditions see, e.g., Ikeda and Watanabe (1981). On a connection between strong and weak solutions, see, e.g., Kabanov and Pergamenshchikov (1991).)

For each \( T > 0 \) we define a (cost) functional

\[
J_{xT}(u) = \int_0^T q(x(t), u(t))dt,
\]

(1.2.14)

where \( q : \mathbb{R}^l \times \mathbb{R}^m \to \mathbb{R} \).

The definition of a strategy in continuous time requires some details which we relegate to Section 2.1. For while we take this and other quite understandable counterparts of the definitions of the discrete time framework for granted, using similar notations for the process and the policy generated by a strategy \( \psi \).

Continuous-time counterparts for \( U_{\psi, x}, X_{\psi, x} \) from the previous section are denoted by \( U_{\psi, x}(t), X_{\psi, x}(t) \), respectively. We fix \( \mathcal{U} \), a class of admissible policies \( u(t) \) on \([0, \infty)\).

Not specifying concrete conditions under which it is true, we suppose that for any \( u \in \mathcal{U} \) the map \( J_{xT}(u) \), as a function of \( (\omega, T) \in \Omega \times \mathbb{R} \), is measurable with respect to \( \mathcal{F} \times \mathcal{B} \), where \( \mathcal{B} \) is the Borel \( \sigma \)-algebra. This requirement is that on \( \mathcal{U} \).

Assume that for each \( T \) there exists a strategy \( \hat{\psi}^T \) optimal in the mean on \([0, T] \). (We skip the obvious counterpart of the corresponding definition from Section 1.2.1). For each initial state \( x \) we denote by \( T\hat{\psi}_x = T\hat{\psi}_x(\cdot) \) the policy generated by \( \hat{\psi}^T \).

Similar to Definition 1, given an initial state \( x \), a policy \( u^* \in \mathcal{U} \) is said to be asymptotically optimal almost surely if for any \( u \in \mathcal{U} \)

\[
\liminf_{T \to \infty} T^{-1}(J_{xT}(u) - J_{xT}(u^*)) \geq 0 \quad \text{a.s.}
\]

(1.2.15)

(The r.-h.s. of (1.2.15) is measurable due to the completeness of \( (\Omega, \mathcal{F}, P) \) and the above measurability assumption.)

Let \( |\cdot| \) denote the Euclidean norm.

**Proposition 3** Suppose the functions \( q \) and \( \sigma \) are bounded, and there exists a constant \( C \) with the following property:

\[
\liminf_{T \to \infty} T^{-1}(J_{xT}(u) - J_{xT}(u^*)) \geq 0 \quad \text{a.s.}
\]

8
For any \(T\) and any pair of states \(z\) and \(z'\), there exist an admissible strategy \(\phi\) and a r.v. \(N\) (perhaps depending on \(T\), \(z\), and \(z'\)) such that

\[
E\{N\} \leq C, \tag{1.2.16}
\]

and

\[
|z - z'|N \leq t \leq T \Rightarrow q(X_{\phi_{z}}(t), U_{\phi_{z}}(t)) = q(\tau \hat{X}_{\phi_{z}}(t), \tau \hat{U}_{\phi_{z}}(t)) \quad \text{a.s.} \tag{1.2.17}
\]

Let for a given \(x\) there exist a policy \(\tilde{u} \in \mathcal{U}\) (perhaps depending on \(x\)), and such that

\[
T^{-1}(J_{\tilde{u}}(\tilde{u}) - J_{\tilde{u}}(\tau \hat{U}_{\phi})) \to 0 \quad \text{a.s., as } T \to \infty.
\]

Then the policy \(\tilde{u}\) is asymptotically optimal almost surely for the \(x\) mentioned.

Note that the main condition (1.2.17), unlike (1.2.8) in the discrete-time case, is infinitesimal in character: the transition time should vanish proportionally to the distance between the initial states. Consider

EXAMPLE 3 which is similar to Example 2 and deals with the linear system

\[
dx(t) = (Ax(t) + Bu(t))dt + \sigma dw(t), \tag{1.2.18}
\]

where \(A, B, \sigma\) are constant matrices of proper dimensions. The cost function \(q\) is still an arbitrary bounded function. As in Example 2, set

\[
y(t) = T \hat{X}_{\phi}(t) - X_{\phi}(t), \quad V(t) = T \hat{U}_{\phi}(t) - U_{\phi}(t), \quad t \in [0, T],
\]

where \(\phi\) is a strategy. Clearly,

\[
dy(t) = (Ay(t) + BV(t))dt, \tag{1.2.20}
\]

with \(y(0) = z - z'\). It is known (see, e.g., again Kwakernaak and Sivan (1972) ) that, if the pair \((A, B)\) is controllable, then the system (1.2.20) can be transferred from the initial state to any finite state in a finite time which may be made arbitrary small. More precisely, let \(\Phi(s, \tau) = e^{A(s-\tau)}\), \(W(\theta) = \int_{0}^{\theta} \Phi(\theta, \tau)BB'\Phi'(\theta, \tau)d\tau\). Then the policy

\[
V_{\theta}(t) = -B'\Phi'(\theta, t)W^{-1}(0, \theta)\Phi(0, 0)(z - z'), \quad \text{for } t \leq \theta, \quad \text{and}
\]

\[
V_{\theta}(t) = 0, \quad \text{for } \theta < t \leq T.
\]

transfers the system in time \(\theta\) from the state \(z - z'\) to the null state. If \(\theta\) is fixed, \(V_{\theta}(t)\) is not random, and consequently, \(y_{\theta}(t)\), the solution to (1.2.20) for \(V(\cdot) = V_{\theta}(\cdot)\), is also non-random.

Let \(\Phi_{T}(t, x)\) be a Markov strategy optimal in the mean on \([0, T]\). Then, as is easy to see from (1.2.19), the Markov strategy \(f_{T}(t, x) = \Phi_{T}(t, x + y(t)) - V_{\mid z - z'}(t)\) transfers the system (1.2.18) from the state \(z\) at the null time moment to the state \(T \hat{X}_{\phi}(\mid z - z'\mid)\) at the moment \(t = \mid z - z'\mid\), if \(\mid z - z'\mid \leq T\). Consequently, condition (1.2.17) holds for \(N = 1\).
1.3 Optimality in Probability.

We turn to a weaker definition of optimality, which allows to weaken conditions (1.2.7), (1.2.16).

**Definition 4** Given an initial state $x$, a policy $\tilde{u} \in \mathcal{U}$ is said to be asymptotically optimal in probability if for any $\varepsilon > 0$, and any policy $u \in \mathcal{U}$

$$P \left( T^{-1} (J_{xT}(\tilde{u}) - J_{xT}(u)) > \varepsilon \right) \to 0, \quad \text{as} \quad T \to \infty. \quad (1.3.1)$$

The definition for the discrete time case is similar.

**Proposition 5** Suppose that the conditions on the strategy $\varphi^n$ and the r.v. $N$ from Proposition 2 hold with the replacement of (1.2.7) by

$$P(N \geq k) = o(k^{-1/2}), \quad \text{as} \quad k \to \infty, \quad \text{uniformly in } z, z' \text{ and } n. \quad (1.3.2)$$

Assume also that for a given $x$ there exists a strategy $\tilde{U}_x \in \mathcal{U}$ and such that for any $\varepsilon > 0$

$$P \left( \frac{1}{n} \left[ J_{xn}(\tilde{U}_x) - J_{xn}(\tilde{U}_x) \right] > \varepsilon \right) \to 0, \quad \text{as} \quad n \to \infty. \quad (1.3.3)$$

Then $\tilde{U}_x$ is asymptotically optimal in probability for the $x$ mentioned.

**Proposition 6** Suppose that the conditions of Proposition 3 hold with the replacement of (1.2.16) by (1.3.2), where now $k$ is any real number, and the relation (1.3.2) is uniform in $z, z'$ and $T$.

Let for a given $x$ there exists a policy $\tilde{u} \in \mathcal{U}$ (perhaps depending on $x$), and such that

$$P \left( \frac{1}{T} \left[ J_{xT}(\tilde{U}_x) - J_{xT}(\tilde{U}_x) \right] > \varepsilon \right) \to 0, \quad \text{as} \quad T \to \infty. \quad (1.3.4)$$

Then the policy $\tilde{u}$ is asymptotically optimal in probability for the $x$ mentioned.

Thus, for optimality in probability we do not need the expectation of the transition time to be finite: this time may be “longer”, and from (1.3.2) it follows that $E\{\sqrt{N}\} \leq k$ constant, would be enough.

It is worthwhile noting also that, when considering optimality in probability, we deal rather not with the asymptotic behavior of the process at infinity but with the probabilities of certain events for a fixed (large) time horizon. Therefore, in this case it would be probably natural to consider not only infinite sequences $u = (u_1, u_2, \ldots)$ but families of policies $\{u^n\} = \{u_{1n}, \ldots, u_{mn}\}$, that is, to work with triangular arrays. It will be done in the next section.

In conclusion note that, when analyzing proofs in Section 3, as well as in many other papers we are referring to, one can notice that, to obtain the same results, it suffices to divide the expression $J_{xT}(u) - J_{xT}(u^*)$ not by $T$ but by some function $r(T) = o(T)$. (In particular, it is connected with the fact that we proceed from a LLN for martingales with finite variances.) The circumstance mentioned was explored in Di Masi and Kabanov (1998), which allowed to come to a definition of the so called $h$-optimality.
As to the order of the function \( r \) needed, complete results were obtained for the classical linear model with quadratic costs in Leizarovitz (1987) (who considered the so-called overtaking optimality), and in Presman (1997) and Belkina, Kabanov, Presman. It was shown that in the model mentioned it is enough to divide by any \( r(T) \to \infty \) for optimality in probability, and by a slowly varying function for optimality a.s.; see the papers mentioned for details. The concrete form of the cost functional in these papers is essential. It is not clear yet to what extent such a result may be true in general.

2  The general framework and main theorems

Generalizations concern the following.

1. We consider more general, in particular, non-homogeneous, processes.
2. Conditions of the type (1.2.8) and (1.2.17) require complete coupling of two certain trajectories. Below we assume these trajectories only to approach each other.
3. Instead of the boundedness of the function \( q \) itself, we require in Theorem 8 below the boundedness of some moment characteristics.
4. The class of admissible policies below may include not only infinite sequences of policies but triangular arrays of those.

General results for the continuous case are somewhat more restrictive than those for the discrete case, and therefore are less complicated. So, we start first with the former case.

2.1 Continuous time

Consider a \( \mathbb{R}^l \)-valued random process \( x(t) \) defined by

\[
 dx(t) = b(x(t), u(t), t)dt + \sigma(x(t), u(t), t)dw(t). \tag{2.1.1}
\]

The processes \( w(\cdot) \), \( u(\cdot) \), and the dimensions of \( b \) and \( \sigma \) are defined as in Section 1.2.2. Set \( x(0) = x \).

A cost functional is

\[
 J_{xT}(u) = \int_0^T g(x(t), u(t), t)dt, \tag{2.1.2}
\]

where \( g : \mathbb{R}^l \times \mathbb{R}^m \times \mathbb{R} \to \mathbb{R} \).

For each \( T \) we denote by \( \mathcal{U}_T \) a fixed class of admissible policies on \([0, T]\). We do not exclude that a process \( u(\cdot) \) from \( \mathcal{U}_T \) depends on \( T \), that is, in general we deal with a family of policies \( u = \{ u_T(t), 0 \leq t \leq T \}, 0 \leq T < \infty \). The same concerns strategies which also may depend on a time horizon.

Set \( \mathcal{U} = \{ \mathcal{U}_T, 0 \leq T < \infty \} \), and \( u = \{ u_T(t), 0 \leq t \leq T \}, 0 \leq T < \infty \), a particular family of policies. We keep the same term policy for families \( u \) too.

The main motivation for such a setup is not that it allows to extend possible ways of control, and in particular, to include into the class of admissible policies those optimal in the mean (though both things are desirable), but in the fact that in this case we establish a stronger optimality property.
Namely we prove that a certain policy \( \tilde{u} \) is the best, in a certain sense, not only with respect to controls generated by policies \( u(t) \) defined for the infinite period \([0, \infty)\), but with respect to controls generated by families of policies. Clearly, it is stronger.

On the other hand, it is worth emphasizing that the triangular-arrays scheme does include the traditional framework when we restrict ourselves to processes \( (\mathcal{U}, \mathcal{T}, \mathbf{P}) \) generated by policies on \([0, \infty)\) (to reduce the former framework to the later it is enough to assume that the class \( \mathcal{U} \) is such that \( u_T(t) = u_T(t) \) for any \( t \in [0, T], T \leq T' \), and \( u = \{u_T(t)\} \in \mathcal{U} \).

So, if the reader are interested only in the traditional framework (or just dislikes families of policies), she/he may have in mind, when reading, just the classical scheme ignoring the second index \( T \) when it appears. It should not cause a misunderstanding. The same concerns the discrete time scheme in the next section.

Below we keep the assumption on the measurability of \( J_{\xi T}(u_T) \) from Section 1.2.2, and sometimes we write \( J_{\xi T}(u) \) instead of \( J_{\xi T}(u_T) \) for \( u_T \in \mathcal{U}_T \).

Assume that for each \( T \) there exists a strategy \( \hat{\psi}^T \) optimal in the mean on \([0, T] \). For each initial state \( x \) we denote by \( T \hat{x} = T \hat{U}_x(t) \) the policy generated by \( \hat{\psi}^T \).

The definition of asymptotic optimality almost surely is the same as in (1.2.15) with the exception that now in general \( u \) and \( u^* \) are families of policies from \( \mathcal{U} \).

Let \( C([0, \infty); \mathbb{R}^d) \) be the space of continuous \( \mathbb{R}^d \)-valued functions on \([0, \infty) \). As usual, we endow this space with the topology of the convergence uniform on compact sets. We consider only admissible feedback policies, that is, random processes \( u(\cdot) \) progressively measurable with respect to the flow generated by the corresponding solutions \( x(\cdot) \) to equation (2.1.1). It presupposes the existence of the corresponding strategy, that is, a measurable map \( \psi : C([0, \infty); \mathbb{R}^d) \times [0, \infty) \rightarrow \mathbb{R}^m \), such that \( u(t) = \psi(x(t), t) \) a.s. for any \( t \geq 0 \), and the r.v. \( \psi(x(\cdot), t) \) is measurable with respect to \( \sigma \{x(s), 0 \leq s \leq t\} \). A policy of this type, and the corresponding process are called the policy and the process, respectively, generated by the strategy \( \psi \).

If a strategy \( \psi(x(t), t) = \phi(x(t), t), t \geq 0 \), for a measurable \( \phi : \mathbb{R}^d \times [0, \infty) \rightarrow \mathbb{R}^m \), then \( \psi \) is a Markov strategy. If \( \phi(x, t) = \phi(x) \), for a \( \phi : \mathbb{R}^d \rightarrow \mathbb{R}^m \), then the strategy is stationary.

We always assume the coefficients \( b \) and \( \sigma \), and any strategy \( \psi \) to be such that there exists a unique strong solution to the equation (2.1.1) with the replacement \( u(t) \) by \( \psi(x(t), t) \). Conditions when it is true are well known; see, e.g., Ikeda and Watanabe (1981), Chapter IV, Theorem 3.1; Karatzas and Shreve (1991), Theorems 5.2.5, 5.2.9. We do not specify such conditions here, but just assume the existence and uniqueness of a strong solution.

The same concerns the regularity of conditional expectations under considerations: we assume that it is always true; see, e.g., Evstigneev (1986).

For a fixed \( T \) consider the value (Bellman) function for the minimization problem concerning the expectation of (2.1.2), that is, the function

\[
V(t, x) = \inf_{u(\cdot)} \left\{ \int_0^T g(x(s), u(s), s) ds \mid x(t) = x \right\}.
\]

For \( r \in \mathbb{R}^m \) denote by \( \mathcal{A}^r \) the operator

\[
\mathcal{A}^r V(t, x) = \frac{1}{2} \sum_{i,j=1}^l a_{ij}(x, r, t) \frac{\partial^2 V}{\partial x_i \partial x_j}(t, x) + \sum_{i=1}^l b_i(x, r, t) \frac{\partial V}{\partial x_i}(t, x),
\]
where the matrix \( \{a_{ij}\} = \sigma \sigma' \). We assume in the sequel that the functions \( b(\cdot, \cdot, \cdot), g(\cdot, \cdot, \cdot) \) are such that:

(a) the Bellman function exists and satisfies the equation

\[
\inf_{r \in \mathbb{R}} \left\{ \frac{\partial V}{\partial t}(t, x) + A'V(t, x) + g(x, r, t) \right\} = 0, \quad V(T, x) \equiv 0; \tag{2.1.3}
\]

(b) there exists a policy optimal in the mean on an interval \([0, T]\), and it is generated by an admissible Markov strategy \( \hat{\phi}_T(x, t) \) which satisfies the equation

\[
\frac{\partial V}{\partial t}(t, x) + A\hat{\phi}_T(x, t)V(t, x) + g(x, \hat{\phi}_T(x, t), t) = 0. \tag{2.1.4}
\]

Sufficient conditions on functions \( b \) and \( g \), under which the last requirements are fulfilled, are also well known; see, e.g., Fleming and Rishel (1975).

For a fixed \( t \), a Markov strategy \( \phi \), and a state \( z \), consider the policy

\[
u_z(s) = \phi(x(s), s), \quad s \geq t, \tag{2.1.5}\]

where \( x(t) \) is a solution to (2.1.1) for \( u(s) = \phi(x(s), s), s \geq t \), with the initial condition \( x(t) = z \). In other words, (2.1.5) is the policy generated by \( \phi \), starting from the moment \( t \), provided that the process at this moment was in the state \( z \).

We denote such a policy by \( U_{\phi z}[s; t] \), and the corresponding process - by \( X_{\phi z}[s; t] \), \( s \geq t \).

If in this case we deal with a strategy \( \hat{\phi} \) optimal in the mean on \([0, T]\), then we will use the notations \( \hat{\tau} \hat{U}_z[s; t], \hat{\tau} \hat{X}_z[s; t] \) omitting the index \( \hat{\phi} \). The processes \( \hat{\tau} \hat{U}_x[s, 0], \hat{\tau} \hat{X}_x[s, 0] \) are denoted by \( \hat{\tau} \hat{U}_x(s), \hat{\tau} \hat{X}_x(s) \), respectively.

We are able now to formulate main conditions.

**Condition A.** There exists a constant \( A \) with the following property:

for any \( T \), any pair of (non-random) states \( z \) and \( z' \), and all \( t \leq T \) there is a Markov strategy \( \phi \) such that

\[
E \left| \int_t^T \left( g(X_{\phi z}[\tau; t], U_{\phi z}[\tau; t]), \tau \right) - g(\hat{\tau} \hat{X}_{z'}[\tau; t], \hat{\tau} \hat{U}_{z'}[\tau; t], \tau) \right| d\tau \leq A|z - z'|. \tag{2.1.6}\]

Since in general the left hand side of (2.1.6) has an order of \( T - t \), inequality (2.1.6) may hold if the values of the cost \( g(\cdot, \cdot, \cdot) \) along two trajectories, \( X_{\phi z} \) and \( \hat{\tau} \hat{X}_{z'} \), approach each other for large \( \tau \). So, Condition A is that on the existence of a strategy \( \phi \) for which it is true. The term \( |z - z'| \)

in the right-hand side of (2.1.6) means that we impose a version of the Lipschits condition with respect to initial states.

For optimality in probability it suffices to impose a weaker

**Condition A'.** There exists a function \( \delta(T) \to 0 \), as \( T \to \infty \), such that for any \( T \), any pair of states \( z, z' \), and all \( t \leq T \) there is an admissible Markov strategy \( \phi \) such that

\[
E \left| \int_t^T \left( g(X_{\phi z}[\tau; t], U_{\phi z}[\tau; t]), \tau \right) - g(\hat{\tau} \hat{X}_{z'}[\tau; t], \hat{\tau} \hat{U}_{z'}[\tau; t], \tau) \right| d\tau \leq \delta(T)T^{1/2}|z - z'|. \tag{2.1.7}\]
**Theorem 7** Let the matrix-function $\sigma(\cdot)$ be bounded.

(1) If Condition A' holds, then for any initial state $x$ the family of policies $\{_{T}^{\hat{U}}x\}$, $0 \leq T < \infty$, is asymptotically optimal in probability.

(2) Suppose Condition A holds, and the function $g$ is bounded from below. Let for a given $x$ and a policy $u^* \in \mathcal{U}$

$$\lim_{T \to \infty} T^{-1} (J_{xT}(u^*) - J_{x[T]}([T]^{\hat{U}}x)) = 0 \quad a.s. \tag{2.1.8}$$

Then $u^*$ is asymptotically optimal a.s. for $x$ mentioned.

First, we comment (2.1.8) which appears only because we consider triangular arrays. This condition means, in particular, that in order to claim that $T^{\hat{U}}x$ is optimal a.s., we should verify that

$$T^{-1} (J_{xT}(\bar{u}) - J_{x[T]}([T]^{\bar{U}}x)) \to 0 \quad a.s., \quad T \to \infty. \tag{2.1.9}$$

It is not difficult to check that the boundedness of $q$ is not sufficient for (2.1.9) to hold automatically. Conditions (2.1.8) or (2.1.9) look very weak, but in the general case we could not avoid it completely.

On the other hand, if there exists a policy $\bar{u}(t)$ on $[0, \infty)$ for which

$$T^{-1} (J_{xT}(\bar{u}) - J_{x[T]}([T]^{\bar{U}}x)) \to 0 \quad a.s., \quad T \to \infty, \tag{2.1.10}$$

then (2.1.9) is true at least for bounded $g$'s. In this case both policies, $\bar{u}$ and $T^{\hat{U}}x$, are asymptotically optimal a.s. (For $\bar{u}$ to be optimal, it suffices to require (2.1.10) to be true only for integer $T$'s.)

Thus, in the traditional framework of policies $u(t)$ on $[0, \infty)$ from a class $\mathcal{U}_\infty$, if (2.1.10) holds, the condition (2.1.9) is redundant.

### 2.2 Discrete time

The general scheme of stochastic control in discrete time may look as follows.

Let $\{\Omega, \mathcal{F}, \mathcal{F}', \mathcal{P}\}$ be a probability space with a flow of $\sigma$-algebras $\{\mathcal{F}' \subset \mathcal{F}\}_{t=0}^\infty$, and $\{(A_t, \mathcal{A}_t)\}_{t=1}^\infty$ be a sequence of measurable spaces of a general nature.

The flow $\{\mathcal{F}'\}$ corresponds to the process under consideration: we do not need to introduce this process explicitly. The $\sigma$-algebra $\mathcal{F}'$ “corresponds to possible initial states” of the process. If $\mathcal{F}'$ is trivial, we interpret it as that the initial state is fixed.

The space $(A_t, \mathcal{A}_t)$ is interpreted as that of actions made at time $t$.

Set $A' = A_1 \times \cdots \times A_t$, $\mathcal{A}' = \mathcal{A}_1 \times \cdots \times \mathcal{A}_t$, $d' = (a_1, \ldots, a_t)$ for $a_j \in A_j$.

The next primitive is a sequence of additive cost functionals

$$J_n(\omega, d^n) = \sum_{i=1}^n g_i(\omega, d^i), \tag{2.2.1}$$

where a function $g_i(\omega, d^i)$ is $\mathcal{F}' \times \mathcal{A}'$-measurable.

(In general, one may endow $g_i$ with the second index $n$ allowing $g_i$ to depend on the time horizon. We skip it here to avoid too cumbersome formulas, though in some sense it would be even more convenient to write $g_{ln}$ instead of $g_l$. In this case one may assume that a normalization
of the sum (say, by dividing by \( n \)) have been already done. (For example, instead of writing
\[
\frac{1}{n}\sum_{t=1}^{n} g_t(\omega, a') \quad \text{one can write } \sum_{t=1}^{n} g_t(\omega, a') \quad \text{setting } g_t(\omega, a') = g_t(\omega, a')/n.
\] Nevertheless, in this paper we stick with the traditional way of writing to make the order of some asymptotic relations below more explicit.)

We will write sometimes \( g_t(\omega, a^n) \) instead of \( g_t(\omega, a') \), keeping in mind that \( g_t \) depends only on the first \( t \) coordinates of the vector \( a^n \).

For a fixed \( n \), a policy is a vector \( u^n = (u_{1n}, \ldots, u_{nn}) \), where \( u_{tn} : \Omega \to \mathcal{A}_t \) is \( \mathcal{F}^{t-1} \)-measurable, so the control may depend on the time horizon. Sometimes, when it does not cause a misunderstanding, we skip the index \( n \). The same concerns strategies below.

Let \( \mathcal{U}_n \) be the class of admissible policies for a fixed \( n \), \( \mathcal{U} = \{ \mathcal{U}_n \} \), and \( u = \{ u^n \} \) where \( u^n \in \mathcal{U}_n \). In general \( u \) is a triangular array. To reduce this scheme to the traditional framework, we should assume that for each \( u = \{ u^n \} \in \mathcal{U} \), it is true that \( u_{tn} = u_{tn'} \) for \( t \leq n \leq n' \). As in the previous section the reader, if she/he wishes, can think just about the traditional scheme ignoring the second index \( n \).

Using, as above, Greek letters for strategies, we define for a fixed \( n \) a strategy as a vector \( \varphi^n = \{ \varphi_1, \ldots, \varphi_n \} \), where \( \varphi_1 : \Omega \to \mathcal{A}_1 \), and \( \varphi_t : \Omega \times \mathcal{A}_t \rightarrow A_t \) is \( \mathcal{F}^{t-1} \times \mathcal{A}_t \)-measurable for \( t = 2, \ldots, n \).

We illustrate the above scheme by the example of the Markov chain from Section 1.2.1. Let the r.v.’s \( \xi_t \) from (1.2.1) be defined on a space \( \{ \Omega, \mathcal{F}, \mathbb{P} \} \), \( \mathcal{F}^0 \) be trivial, and \( \mathcal{F}^t = \sigma(\xi_t) \). In view of (1.2.1), there exists a \( \mathcal{F}^t \times \mathcal{A}_t \)-measurable function \( g_t(\omega, a_t) = q(x_t, a_t) \). For any strategy \( \psi_t(x^{t-1}) \) from Section 1.2.1 there exists a \( \mathcal{F}^{t-1} \times \mathcal{A}_t \)-measurable function \( \varphi_t(\omega, a^{t-1}) = \psi_t(x^{t-1}) \). (Both functions \( g \) and \( h \) may depend on the fixed initial state \( x \).) So, we are in the frame of the above scheme.

The above general scheme allows to avoid some superfluous considerations, and makes proofs simpler. Certainly, it does not mean that a knowledge about the concrete structure of the process (say, the representations (1.2.1) and (1.2.2)) cannot be useful for other purposes. Note also that an additional information on the process and on the cost functional may be described in terms of the \( \sigma \)-algebras with respect to which the functions \( g_t \) are measurable. In particular models these \( \sigma \)-algebras are, as a rule, simpler than \( \mathcal{F}^t \times \mathcal{A}_t \) (see, e.g., Rotar (1986a)).

Furthermore, for any strategy \( \varphi^n \) we define a set of operators \( \mathcal{F}^n, t = 1, 2, \ldots, n \); transforming the space of policies into itself, and such that the policy \( \mathcal{F}^n u^n \) coincides with \( u^n \) up to the moment \( t \), and starting with that moment, \( \mathcal{F}^n u^n \) governs the process in accordance with the strategy \( \varphi^n \). More precisely, for \( \tilde{u}^n = (\tilde{u}_1, \ldots, \tilde{u}_n) = \mathcal{F}^n u^n \),

\[
\tilde{u}_j = \begin{cases} 
    u_j & \text{if } j < t; \\
    \varphi_j(\omega, \tilde{u}_1(\omega), \ldots, \tilde{u}_{j-1}(\omega)) & \text{if } j \geq t.
\end{cases}
\]

The policy \( \mathcal{F}^n u^n \) is the policy generated by the strategy \( \varphi^n \) from the initial moment, so it does not depend on \( u^n \). By convention, \( \mathcal{F}^n u^n = u^n \) for \( t > n \).

Set \( E_t(\cdot) = E(\cdot | \mathcal{F}^{t-1}) \), \( E_0 = E \).

In the non-homogeneous case we need a bit stronger optimality than optimality in the mean. Namely, we assume that there exists a persistently optimal strategy \( \hat{\varphi}^n \), that is, a strategy such that for all \( u^n \in \mathcal{U}_n \) and \( t = 0, \ldots, n \)

\[
E_t J_n(\omega, \mathcal{F}_t \hat{\varphi}^n u^n) \geq E_t J_n(\omega, \mathcal{F}_t u^n) \quad \text{a.s.,}
\] (2.2.2)
where the operators $\hat{\Phi}^n$ correspond to $\hat{\Phi}^n$. (For persistent optimality see, e.g., Kertz (1982)). Set $\hat{\mu}^n = \hat{\Phi}^n u^n$, and $\hat{u}^n = \{\hat{\mu}^n\}$.

**Condition B.** There exist $p > 2$ and a constant $B_p$ with the following property:
for any $n$, all pairs of policies $u^n$, $\tilde{u} \in \mathcal{U}_n$, and all $t \leq n$ there is a policy $z^n \in \mathcal{U}_n$ such that

$$z^n = (\tilde{u}, \cdots, \tilde{u}, z_{t+1}, \cdots, z_n),$$

(2.2.3)

and

$$E \left| \sum_{j=t+1}^{n} \{g_j(\omega, z^n) - g_j(\omega, t+1 \hat{\Phi}^n u^n)\} \right|^p \leq B_p,$$

(2.2.4)

The last condition may be interpreted in the same way as we interpreted Condition A in Section 2.1.

**Condition D.** There exists a constant $D$ such that for any $n$, any pair of policies $u^n$, $\tilde{u} \in \mathcal{U}_n$, and all $t \leq n$, there is a policy (2.2.3) such that

$$E \left[ \sum_{j=t+1}^{n} \{g_j(\omega, z^n) - g_j(\omega, t+1 \hat{\Phi}^n u^n)\} \right]^2 \leq \gamma(n)n.$$

(2.2.6)

**Condition B’.** There exists a function $\gamma(n) \to 0$, as $n \to \infty$, with the following property:
for all $n$, all pairs $u^n$, $\tilde{u} \in \mathcal{U}_n$, and all $t \leq n$ there is a policy $z^n \in \mathcal{U}_n$ such that (2.2.3) is true, and

$$E \left[ \sum_{j=t+1}^{n} \{g_j(\omega, z^n) - g_j(\omega, t+1 \hat{\Phi}^n u^n)\} \right]^2 \leq \gamma(n)n.$$

(2.2.6)

**Condition D’.** Condition B’ is true with replacement (2.2.6) by

$$E_{t+1} \left| \sum_{j=t+1}^{n} \{g_j(\omega, z^n) - g_j(\omega, t+1 \hat{\Phi}^n u^n)\} \right| \leq \gamma(n)\sqrt{n} \quad \text{for all } t = 0, 1, \ldots .$$

(2.2.7)

**Theorem 8** (1) Let one of conditions B or D hold, and for all $u = \{u^n\} \in \mathcal{U}$

$$(1/n) \sum_{t=1}^{n} \{g_t(\omega, u^n) - E_t g_t(\omega, u^n)\} \to 0, \quad \text{as} \quad n \to \infty,$$

(2.2.8)

almost surely. Then the sequence of policies $\hat{u}$ is asymptotically optimal almost surely.

(2) If one of conditions B’ or D’ holds, and for all $u = \{u^n\} \in \mathcal{U}$ (2.2.8) is true in probability, then the sequence $\hat{u}$ is asymptotically optimal in probability.

The following

**REMARKS** concern sufficient conditions for the validity of (2.2.8).
(1) Since the summands in (2.2.8) are martingale-differences with respect to \( \{ F_t \} \), it is clear that (2.2.8) holds in probability if, say,

\[
\frac{1}{n^2} \sum_{t=1}^{n} E [g_t(\omega, u^n) - E_t g_t(\omega, u^n)]^2 \to 0, \quad n \to \infty.
\] (2.2.9)

This is already is quite mild.

(2) Obviously, (2.2.9) holds if the functions \( g_t(\cdot) \) are uniformly bounded, that is,

\[
\sup_{t, \omega, a'} |g_t(\omega, a')| < \infty.
\] (2.2.10)

(3) Using the Borel-Cantelli lemma, it is not difficult to get (see also arguments in (3.3.6) below) that (2.2.8) is true almost surely if for some \( p \geq 2 \) and all \( u \in \mathcal{U} \)

\[
\sum_{n=1}^{\infty} \frac{1}{n^p} \left[ \sum_{t=1}^{n} \| g_t(\omega, u^n) - E_t g_t(\omega, u^n) \|^2_p \right]^{p/2} < \infty,
\] (2.2.11)

where \( \| \cdot \|_p \) is the norm in \( L_p \).

(4) Clearly, (2.2.11) is fulfilled if, for example, there exist constants \( C > 0 \) and \( p > 2 \) such that for all policies \( u^n \in \mathcal{U}_n \)

\[
\| g_t(\omega, u') - E_t g_t(\omega, u') \|_p \leq C.
\] (2.2.12)

Clearly, (2.2.10) implies (2.2.12).

3 Proofs

3.1 Proofs of Propositions

Proofs of Propositions 3 and 6. We show that these propositions follow from Theorem 7. As was noted after the formulation of theorem 7, in the case under discussion condition (2.1.8) holds automatically.

Since the process is homogeneous, when verifying Conditions A or A', it suffices to set in (2.1.6) and (2.1.7) \( t = 0 \). Let \( N \) and \( \varphi \) be the r.v. and the strategy, respectively, defined in the propositions. Then in virtue of (1.2.17) for any initial states \( z, z' \)

\[
\left| \int_0^T \left( g(X_{\varphi}, [\tau, 0]), U_{\varphi} \tau, 0, \tau \right) - g(T \hat{X}_{\varphi} \tau, 0, T \hat{U}_{\varphi} \tau, 0, \tau) \right| d\tau \right| \leq 2NG|z - z'| \quad a.s., \] (3.1.1)

where \( G = \sup \| g \| \).

So, if \( E\{N\} \leq C \), then (2.1.6) holds with \( A = 2GC \).

To derive (2.1.7) from conditions of Proposition 6, we observe that, if \( N > T \), then the l.-h.s. of (1.2.17) is an empty set, and consequently, (1.2.17) holds automatically. Hence, condition (1.2.17) remains to be true if we replace \( N \) by its truncation \( N(T) = N \), if \( N \leq T \), and \( N(T) = T \), if \( N \geq T \). It means that in the r.-h.s. of (3.1.1) we can replace \( N \) by \( N(T) \).
In view of (1.3.2), there exists a function $\delta_0(t) \to 0$, as $t \to \infty$, which does not depend on $z, z'$ and $T$, and is such that $P(N > t) \leq \delta_0(t)\sqrt{T}$. Hence

$$E\{N(T)\} = E\{N; N \leq T\} + TP(N > T) \leq E\{N; N \leq T\} + \delta_0(T)\sqrt{T}.$$  

It is easy to show that (1.3.2) implies $E\{N; N \leq T\} \leq \delta_1(T)\sqrt{T}$, where a function $\delta_1(T) \to 0$, and is determined only by the function $\delta_0(T)$.

Hence, (1.2.17) implies (2.1.7) with $\delta(T) = 2G(\delta_1(T) + \delta_0(T))$. ■

**Proof of Propositions 2 and 5.** We show that these propositions follow from Theorem 8. We have assumed the existence of strategies $\hat{\Phi}^n$ optimal in the mean on $\{0, n\}$, that is, strategies which generates optimal in the mean policies for all initial states. Since the scheme is homogeneous, it means that $\hat{\Phi}^n$ are persistently optimal. So, we can fix an initial state $x$, and consider the problem in the frame of the above general scheme, assuming $\mathcal{F}^0$ to be trivial.

Since $g$ is bounded, (2.2.8) in this case holds automatically (see also remarks which follow Theorem 8). Since the scheme is homogeneous and $\mathcal{F}^0$ is trivial, it suffices to verify (2.2.5) for $t = 0$ and replace the conditional expectation by the unconditional. The derivation of (2.2.5) and (2.2.7), that is, Conditions $D$ and $D'$, from (1.2.7) and (1.3.2), respectively, is similar to the arguments in the previous subsection. ■

### 3.2 Proofs for the continuous case

#### 3.2.1 The main inequality

We start with a multidimensional version of Theorem 2.1 from Presman, Rotar, Taksar (1993). For brevity, we denote here the optimal in the mean policy and the process it generates by $\hat{u}_T(t)$ and $\hat{x}_T(t)$, respectively, not specifying an initial state.

**Lemma 9** For an arbitrary policy $u$ and the process $x(t)$ this policy generates,

$$J_{xT}(u) - J_{xT}(\hat{u}_T) \geq \int_0^T \left[ \frac{\partial V}{\partial x}(t, x(t))\sigma(x(t), u(t), t) - \frac{\partial V}{\partial x}(t, \hat{x}_T(t))\sigma(\hat{x}_T(t), \hat{u}_T(t), t) \right] dt \quad (3.2.1)$$

(where $\partial V/\partial x$ is a vector $(\partial V/\partial x_1, ..., \partial V/\partial x_l)$, and $\partial V/\partial x \sigma dw$ is the dot product $\frac{\partial V}{\partial x} \cdot \sigma dw$).

**PROOF** is similar to that from Presman, Rotar, Taksar (1993). By (2.1.1) and Ito’s formula,

$$V(T, x(T)) - V(0, x) = \int_0^T \frac{\partial V}{\partial t}(t, x(t)) dt$$

$$+ \frac{1}{2} \sum_{i,j=1}^l a_{ij}(x(t), u(t), t) \frac{\partial^2 V}{\partial x_i \partial x_j}(t, x(t)) + \sum_{i=1}^l b_i(x(t), u(t), t) \frac{\partial V}{\partial x_i}(t, x(t)) dt$$

$$+ \int_0^T \sum_{i=1}^l \frac{\partial V}{\partial x_i}(t, x(t)) \sum_{j=1}^d \sigma_{ij}(x(t)), u(t), t) dw_j(t). \quad (3.2.2)$$
In view of (2.1.3), the integrand in the first integral in (3.2.2) is not less than \(-g(x(t), u(t), t)\). In view of the second relation in (2.1.3), the l.-h.s. in (3.2.2) is equal to \(-V(0, x)\). Thus,

\[ -V(0, x) \geq - \int_0^T g(x(t), u(t), t) dt + \int_0^T \frac{\partial V}{\partial x}(t, x(t)) \sigma(x(t), u(t), t) dw(t). \]

Similarly, (2.1.4) and (3.2.2) imply that

\[ -V(0, x) = - \int_0^T g(\hat{x}_T(t), \hat{u}_T(t), t) dt + \int_0^T \frac{\partial V}{\partial x}(t, \hat{x}_T(t)) \sigma(\hat{x}_T(t), \hat{u}_T(t), t) dw(t). \]

The last two relations imply (3.2.1).

### 3.2.2 Proof of Theorem 7

Let \( u \) be a policy, and \( x(t) \) be the process generated by \( u \). Making use of the boundedness of the function \( \sigma \), (3.2.1), and Doob’s inequality for locally square-integrable martingales (see, e.g., Liptser and Shiryaev (1972)), we have

\[
P\left( (J_T(\hat{u}_T) - J_T(u_T)) > \varepsilon T \right) 
\leq P \left( \int_0^T \left( \frac{\partial V}{\partial x}(t, \hat{x}_T(t)) \sigma(\hat{x}_T(t), \hat{u}_T(t), t) - \frac{\partial V}{\partial x}(t, x_T(t)) \sigma(x(t), u(t), t) \right) dw(t) > \varepsilon T \right) 
\leq \frac{l \sup ||\sigma||^2}{\varepsilon^2 T^2} \mathbb{E} \left\{ \int_0^T \left[ \left| \frac{\partial V}{\partial x}(t, \hat{x}_T(t)) \right|^2 + \left| \frac{\partial V}{\partial x}(t, x_T(t)) \right|^2 \right] dt \right\},
\]

(3.2.3)

where \( ||\cdot|| \) is the Euclidean norm, \( ||\cdot|| \) is the matrix norm.

Below we omit, for brevity, index \( T \) in notations of policies and processes.

For \( j = 1, \ldots, l \), set \( \Delta_j = (0, 0, \ldots, 0, \Delta, 0, \ldots, 0) \), where the number \( \Delta \) is the \( j \)-th element of the vector \( \Delta_j \).

For any fixed \( x \), any \( t \in [0, T] \), and an arbitrary admissible strategy \( \varphi \),

\[
V(t, x) \leq E \left( \int_t^T g(X_{\varphi \tau}[\tau, t], U_{\varphi \tau}[\tau, t], \tau) d\tau \right) 
= V(t, x + \Delta_j) + E \left( \int_t^T g(X_{\varphi \tau}[\tau, t], U_{\varphi \tau}[\tau, t], \tau) d\tau \right) 
- E \left( \int_t^T g(\hat{X}_{x+\Delta_j}[\tau, t], \hat{U}_{x+\Delta_j}[\tau, t], \tau) d\tau \right).
\]

Consequently,

\[
V(t, x) - V(t, x + \Delta_j) 
\leq \left| E \int_t^T \{ g(X_{\varphi \tau}[\tau, t], U_{\varphi \tau}[\tau, t], \tau) d\tau - g(\hat{X}_{x+\Delta_j}[\tau, t], \hat{U}_{x+\Delta_j}[\tau, t], \tau) \} d\tau \right|. \quad (3.2.4)
\]
Replacing in (3.2.4) \( \varphi \) by another strategy \( \tilde{\varphi} \), and interchanging \( x \) and \( x + \Delta j \), one can easily get that

\[
V(t, x) - V(t, x + \Delta_j) \\
\geq - \left| E \int_t^T \left\{ g \left( X_\varphi(x+\Delta_j)[\tau, t], U_\varphi(x+\Delta_j)[\tau, t], \tau \right) - g \left( \hat{X}_x(\tau, t], \hat{U}_x(\tau, t], \tau \right) \right\} d\tau \right|.
\]

Let now strategies \( \varphi \) and \( \tilde{\varphi} \) be such that the policies they generate posses Property A', where the role of the pair \((z, z')\) is played by \((x, x + \Delta_j)\) and \((x + \Delta_j, x)\), respectively. Then from (3.2.4) and (3.2.5) it follows that

\[
\left| V(t, x) - V(t, x + \Delta_j) \right| \\
\leq \left| E \left\{ \int_t^T g \left( X_\varphi(\tau, t], U_\varphi(\tau, t], \tau \right) - g \left( \hat{X}_x(\tau, t], \hat{U}_x(\tau, t], \tau \right) \right\} d\tau \right| + \left| E \left\{ \int_t^T g \left( X_\varphi(x+\Delta_j(\tau, t], U_\varphi(x+\Delta_j)[\tau, t], \tau \right) - g \left( \hat{X}_x(\tau, t], \hat{U}_x(\tau, t], \tau \right) \right\} d\tau \right| \\
\leq 2\delta(T)T^{1} |\Delta_j|
\]

for any \( x, \Delta \) and \( j \). Hence

\[
\left| \frac{\partial V}{\partial x}(t, x) \right|^2 \leq 4l^2\delta^2(T)T
\]

for any (vector) \( x \). Consequently the same inequality is true for any \( x = x(t) \) a.s..(Note also that above \( \partial V/\partial x \) is a vector, and \( |\cdot| \) is the Euclidean norm.)

So, making use of (3.2.3), for any family of policies \( \{u_T(t) \in \mathcal{U} \}, 0 \leq T < \infty \}, \) we have

\[
P \left( (J_{xT}(\hat{u}_T) - J_{xT}(u_T)) > \varepsilon T \right) \leq 4l^2 \sup \|\sigma(.)\|^2 (\varepsilon T)^{-2} \int_0^T \delta^2(T)T dt \\
= 4l^2 \sup \|\sigma(.)\|^2 \varepsilon^{-2} \delta^2(T) \to 0, \quad \text{as} \quad T \to \infty.
\]

The first assertion of Theorem 7 is proved. We turn to the second.

In view of (2.1.8) it suffices to show that for any policy \( u \in \mathcal{U} \),

\[
\limsup_{T \to \infty} T^{-1} \left[ (J_{xT}[\hat{u}_T]) - J_{xT}(u) \right] \leq 0 \quad \text{a.s.}
\]

On the other hand,

\[
P \left\{ \sup_{T > s} \left( T^{-1} (J_{xT}[\hat{u}_T]) - J_{xT}(u) \right) > \varepsilon \right\} \leq \sum_{n=\lfloor s \rfloor}^\infty P \left\{ \sup_{n \leq T < n+1} \left( (J_{xT}(\hat{u}_n) - J_{xT}(u)) > \varepsilon n \right) \right\} \\
\leq \sum_{n=\lfloor s \rfloor}^\infty P \left( (J_{xT}(\hat{u}_n) - J_{xT}(u)) + a > \varepsilon n \right),
\]

where \( a = |\inf g| \).
From this, (3.2.1) and Markov’s inequality, it easily follows that it suffices to show the convergence of the series

$$
\sum_{n=1}^{\infty} \frac{1}{n^p} E \left( \int_{0}^{n} \left[ \frac{\partial V}{\partial x}(t, \hat{x}(t)) \sigma(\hat{u}(t), \hat{u}(t), t) - \frac{\partial V}{\partial x}(t, x(t)) \sigma(x(t), u(t), t) \right] dw(t) \right)^p
$$

(3.2.6)

for some $p > 2$.

Let $\| \cdot \|_p$ be the norm in $L_p$, and $1 = (1, \ldots, 1)$. Denote by $C(\cdot)$ any constant depending only on arguments in (\cdot), and perhaps different in different formulas.

By the Burkholder-Gandhy inequality (see, e.g., Liptser and Shiryaev (1972)), for a constant $C(p)$ depending only on $p$

$$
E \left( \int_{0}^{n} \left[ \frac{\partial V}{\partial x}(t, \hat{x}(t)) \sigma(\hat{u}(t), \hat{u}(t), t) - \frac{\partial V}{\partial x}(t, x(t)) \sigma(x(t), u(t), t) \right] dw(t) \right)^p
\leq C(p) E \left( \int_{0}^{n} \left( \frac{\partial V}{\partial x}(t, \hat{x}(t)) \sigma(\hat{u}(t), \hat{u}(t), t) 1 - \frac{\partial V}{\partial x}(t, x(t)) \sigma(x(t), u(t), t) 1 \right)^2 dt \right)^{p/2}
\leq 2C(p) l \sup \| \sigma(.) \|^2 \left( \int_{0}^{n} \left[ \frac{\partial V}{\partial x}(t, \hat{x}(t)) \right]^2 \left[ \frac{\partial V}{\partial x}(t, x(t)) \right]^2 dt \right)^{p/2}
\leq 2C(p) l \sup \| \sigma(.) \|^2 \left( \left( \int_{0}^{n} \left[ \frac{\partial V}{\partial x}(t, \hat{x}(t)) \right]^2 dt \right)^{p/2} + \left( \int_{0}^{n} \left[ \frac{\partial V}{\partial x}(t, x(t)) \right]^2 dt \right)^{p/2} \right)
\leq C(p, l, \sup \| \sigma(.) \|) \left\{ \left( \int_{0}^{n} \left[ \frac{\partial V}{\partial x}(t, \hat{x}(t)) \right]^2 dt \right)^{p/2} + \left( \int_{0}^{n} \left[ \frac{\partial V}{\partial x}(t, x(t)) \right]^2 dt \right)^{p/2} \right\}.
$$

(Above for a random vector $X$, the symbol $\| X \|_p$ denotes $(E|X|^p)^{1/p}$, where $| \cdot |$ is the Euclidean norm.)

Furthermore, following arguments similar to those from the first part of the proof, it is easy to prove that under (2.1.6)

$$
\left| \frac{\partial V}{\partial x}(t,x) \right|^2 \leq 4lA^2,
$$

and, consequently,

$$
E \left( \int_{0}^{n} \left[ \frac{\partial V}{\partial x}(t, \hat{x}(t)) \sigma(\hat{u}(t), \hat{u}(t), t) - \frac{\partial V}{\partial x}(t, x(t)) \sigma(x(t), u(t), t) \right] dw(t) \right)^p \leq C(p, l, \sup \| \sigma(.) \|) l n^{p/2}.
$$

So, for $p > 2$ the series (3.2.6) converges. Theorem 1 is proved.
3.3 Proofs for the discrete case

We use some facts from Asriev and Rotar (1990), and Rotar (1991). Consider first the strategy \( \hat{\Phi}^n \), the policy \( \hat{u}^n \), and the map \( \hat{\Phi}^n \), defined in Section 2.2. Set \( \hat{u} = \{ \hat{u}^n \} \in \mathcal{U} \), and fix another sequence of policies \( u = \{ u^n \} \).

Below we omit the index \( n \) in \( u_{tn}, \hat{u}_{tn}, \Phi_{tn} \) and \( \hat{\Phi}^n \).

Let r.v.

\[
Y_t = Y_{tn} = E_t \{ J_n(\omega, t_{n+1} \hat{\Phi} u^n) \} - E_t \{ J_n(\omega, t \hat{\Phi} u^n) \}.
\]

By “telescoping”,

\[
J_n(\omega, u^n) = \sum_{t=0}^{n} (E_{t+1} - E_t) \{ J_n(\omega, t_{n+1} \Phi u^n) \} + \sum_{t=1}^{n} Y_t + E \{ J_n(\omega, \hat{u}^n) \}.
\]

In particular case when \( u = \hat{u} \), we have \( Y_t = 0 \text{ a.s.} \), and

\[
J_n(\omega, \hat{u}^n) = \sum_{t=0}^{n} (E_{t+1} - E_t) \{ J_n(\omega, \hat{u}^n) \} + \sum_{t=1}^{n} Y_t + E \{ J_n(\omega, \hat{u}^n) \}.
\]

Subtracting the latter identity from the former, we have (see also Asriev and Rotar (1990))

\[
J_n(\omega, u^n) - J_n(\omega, \hat{u}^n) = \sum_{t=1}^{n} Y_t + \sum_{t=1}^{n} W_t, \tag{3.3.1}
\]

where

\[
W_t = W_{tn} = \sum_{j=t}^{n} [Z_{tj} - E_t Z_{tj}], \quad \text{and} \quad Z_{tt} = Z_{tt}(n) = g_t(\omega, u^n) - g_t(\omega, \hat{u}^n),
\]

\[
Z_{tj} = Z_{tj}(n) = E_{t+1} \{ g_j(\omega, t_{n+1} \hat{\Phi} u^n) - g_j(\omega, \hat{u}^n) \} \quad \text{for} \quad j > t.
\]

(It is worthwhile to recall here that \( g_t(\omega, u^n) \) depends only on the first \( j \) coordinates of the vector \( u^n = (u_1, \ldots, u_n) \).

Note that, taking expectations of (3.3.1), we would come to

\[
E \{ J_n(\omega, u^n) \} - E \{ J_n(\omega, \hat{u}^n) \} = \sum_{t=1}^{n} E \{ Y_t \},
\]

which is similar to the formula for the “strategy defect” from Yushkevich and Chitashvili (1982). The representation (3.3.1) takes into account the “random part of the defect” too.

Since the strategy \( \hat{\Phi}^n \) is persistently optimal, \( Y_t \geq 0 \text{ a.s.} \) for all \( t \leq n \) (see (2.2.2)). Consequently, it remains to prove that

\[
\frac{1}{n} \sum_{t=1}^{n} W_t \to 0, \quad \text{as} \quad n \to \infty, \tag{3.3.2}
\]

almost surely or in probability, depending on what type of optimality we are considering.
Let
\[ V_t = V_{tn} = \sum_{j=t+1}^{n} E_{t+1} \{ g_j(\omega, \hat{u}^n) \}, \quad \hat{V}_t = V_{tn} = \sum_{j=t+1}^{n} E_{t+1} \{ g_j(\omega, {\hat{u}}^n) \}, \quad R_t = R_{tn} = V_{tn} - \hat{V}_{tn}. \]

Then
\[ W_t = Z_{tt} - E_t Z_{tt} + R_t - E_t R_t. \tag{3.3.3} \]

In view of condition (2.2.8)
\[ (1/n) \sum_{i=1}^{n} [Z_{ti} - E_t Z_{tt}] \to 0, \quad \text{as} \quad n \to \infty, \tag{3.3.4} \]
in probability or a.s. (see the formulation of Theorem 8). Thus, it suffices to show that
\[ (1/n) \sum_{i=1}^{n-1} \{ R_i - E_t R_t \} \to 0, \quad \text{as} \quad n \to \infty, \tag{3.3.5} \]
again in probability or a.s.

We fix some \( p > 2 \), and set \( U_t = R_t - E_t R_t \). The summands in (3.3.5) are martingale-differences with respect to \( \{ J_i \} \), and in view of Burkholder’s inequality, there exists a constant \( C_p \) depending only on \( p \) and such that
\[
E \left[ \sum_{t=1}^{n-1} U_t^p \right] \leq C_p E \left[ \sum_{t=1}^{n-1} U_t^2 \right]^{p/2} \leq C_p \left[ \sum_{t=1}^{n-1} \left| U_t \right|^2 \right]^{p/2} \leq 2^n C_p \left[ \sum_{t=1}^{n-1} \left| R_t \right|^2 \right]^{p/2} \tag{3.3.6} \]

By (3.3.6), Markov’s inequality and the Borel-Cantelli lemma, it is easy to show that the almost sure convergence in (3.3.5) takes place if for some \( p > 2 \)
\[
\sum_{n=1}^{\infty} \frac{1}{n^p} \left[ \sum_{t=1}^{n-1} \left| R_t \right|^2 \right]^{p/2} < \infty. \tag{3.3.7} \]

The following reasoning is similar to that in the proof of Theorem 1. We fix \( n \) and \( n \leq t \), and turn to Conditions B or D which deal with a pair of arbitrary policies \( (u^n, \tilde{u}) \). Set first \( \tilde{u} = \hat{u} \). Then the policy (2.2.3) takes the form
\[
z^n = (\hat{u}_1, \cdots, \hat{u}_t, \tilde{z}_{t+1}, \cdots, \tilde{z}_n). \tag{3.3.8} \]

Since the strategy \( \hat{\Phi}^n \) is persistently optimal,
\[
\hat{V}_t \leq E_{t+1} \sum_{j=t+1}^{n} g_j(\omega, z^n) = V_t + E_{t+1} \left( \sum_{j=t+1}^{n} g_j(\omega, \tilde{z}^n) - \sum_{j=t+1}^{n} g_j(\omega_{t+1} \hat{\Phi}^n u^n) \right), \]

23
and

\[ \hat{V}_t - V_t \leq \mathbf{E}_{t+1} \left( \sum_{j=t+1}^{n} g_j(\omega, z^n) - \sum_{j=t+1}^{n} g_j(\omega, t+1 \hat{\Phi}^n u^n) \right). \quad (3.3.9) \]

Now we consider, as the pair \((u^n, t\hat{\Phi}^n u^n)\), the pair \((\hat{u}, t+1 \hat{\Phi}^n u^n)\). Denote the policy (2.2.3) in this case by \(k^n\). Certainly, it may differ from policy (3.3.8). By the definition of the operator \(t+1 \hat{\Phi}^n\), we have \(y = u_j\) for \(j \leq t\), and policy (2.2.3) may be represented in this case as

\[ k^n = \{u_1, \ldots, u_t, k_{t+1}, \ldots, k_n\}. \]

Observe also that \(t+1 \hat{\Phi}^n \hat{u}^n = \hat{u}^n\) for all \(t\). Hence, since the strategy \(\hat{\Phi}^n\) is persistently optimal,

\[ V_t \leq \mathbf{E}_{t+1} \sum_{j=t+1}^{n} g_j(\omega, k^n) = \mathbf{E}_{t+1} \sum_{j=t+1}^{n} g_j(\omega, t+1 \hat{\Phi}^n \hat{u}^n) + \]

\[ + \mathbf{E}_{t+1} \left( \sum_{j=t+1}^{n} g_j(\omega, k^n) - \sum_{j=t+1}^{n} g_j(\omega, t+1 \hat{\Phi}^n \hat{u}^n) \right) = \]

\[ = \mathbf{E}_{t+1} \sum_{j=t+1}^{n} g_j(\omega, \hat{u}^n) + \mathbf{E}_{t+1} \left( \sum_{j=t+1}^{n} g_j(\omega, k^n) - \sum_{j=t+1}^{n} g_j(\omega, t+1 \hat{\Phi}^n \hat{u}^n) \right) = \]

\[ = \hat{V}_t + \mathbf{E}_{t+1} \left( \sum_{j=t+1}^{n} g_j(\omega, k^n) - \sum_{j=t+1}^{n} g_j(\omega, t+1 \hat{\Phi}^n \hat{u}^n) \right), \]

and

\[ V_t - \hat{V}_t \leq \mathbf{E}_{t+1} \left( \sum_{j=t+1}^{n} g_j(\omega, k^n) - \sum_{j=t+1}^{n} g_j(\omega, t+1 \hat{\Phi}^n \hat{u}^n) \right). \quad (3.3.10) \]

Thus, if Condition B holds, then by (3.3.9) and (3.3.10)

\[ \|R_t\|_p = \left\|\hat{V}_t - V_t\right\|_p \leq 2B_p^{1/p}, \quad (3.3.11) \]

and

\[ \sum_{n=1}^{\infty} \frac{1}{n^p} \left[ \sum_{t=1}^{n-1} \|R_t\|_p^2 \right]^{p/2} \leq 2^p B_p \sum_{n=1}^{\infty} \frac{1}{n^{p/2}} < \infty, \quad (3.3.12) \]

if \(p > 2\).

Let now Condition D hold. In this case from (3.3.9), (3.3.10) and (2.2.5) it follows that

\[ |R_t| = |V_t - \hat{V}_t| \leq 2D \quad \text{a.s.,} \]

and the l.-h.s. of (3.3.12) is again finite for \(p > 2\). So, the first assertion of Theorem 8 is proved.

Consider now the convergence in (3.3.5) in probability. Similar to the proof of (3.3.11) one can get that under Condition B'

\[ \mathbf{E} |R_t|^2 \leq 2\gamma(n)n. \]
Then the variance of the left-hand side of (3.3.5) is less than
\[
\frac{1}{n^2} \sum_{t=1}^{n} 2E |R_t|^2 \leq 4\gamma(n) \rightarrow 0. \tag{3.3.13}
\]

In the case of Condition $D'$, we have
\[|R_t| \leq 2\gamma(n)\sqrt{n} \text{ a.s.,}
\]
and (3.3.13) is again true.

Theorem 2 is proved.

References

[1] V.I. Arkin, I.V. Evstigneev (1987), Stochastic models of control and economic

in dynamic control, Stochastics and Stochastic Reports, N 33, pp.1-16.

linear regulator problem with varying parameters, Avtomatika i Telemekhanika, N 2, pp.110-


with a quadratic cost functional, Avtomatika i Telemekhanika, N 3, pp. 106-115 (English

[6] T.A. Konyukhova [Belkina], and V.I. Rotar (1992), Controls asymptotically optimal in prob-
ability and almost surely in the linear regulator problem, Avtomatika i Telemekhanika, N 6,

[7] T.A. Belkina, and V.I. Rotar (1999), On conditions for optimality in probability and almost
surely in a model of controlled diffusion process, Avtomatika i Telemekhanika, N 2, pp. 45-56


Technical, Inc.


[38] V.I.Rotar (1997), Probability Theory, Singapore, New Jersey, etc.