algorithms

MDPI

Article

Optimization of Constrained Stochastic Linear-Quadratic Control on an Infinite Horizon: A Direct-Comparison Based Approach

Ruobing Xue [1], Xiangshen Ye [1,*] and Weiping Wu [2]

[1] Department of Automation, Shanghai Jiaotong University, Shanghai 200240, China; cynthiabaobei@sjtu.edu.cn

[2] School of Economics and Management, Fuzhou University, Fuzhou 350108, China; wu.weiping@fzu.edu.cn

* Correspondence: yesyjs@sjtu.edu.cn

Received: 20 January 2020; Accepted: 19 February 2020; Published: 24 February 2020

Abstract: In this paper we study the optimization of the discrete-time stochastic linear-quadratic (LQ) control problem with conic control constraints on an infinite horizon, considering multiplicative noises. Stochastic control systems can be formulated as Markov Decision Problems (MDPs) with continuous state spaces and therefore we can apply the direct-comparison based optimization approach to solve the problem. We first derive the performance difference formula for the LQ problem by utilizing the state separation property of the system structure. Based on this, we successfully derive the optimality conditions and the stationary optimal feedback control. By introducing the optimization, we establish a general framework for infinite horizon stochastic control problems. The direct-comparison based approach is applicable to both linear and nonlinear systems. Our work provides a new perspective in LQ control problems; based on this approach, learning based algorithms can be developed without identifying all of the system parameters.

Keywords: linear-quadratic; Markov decision process(MDP); conic constraints; stochastic control; direct-comparison based approach

1. Introduction

In this paper we study the discrete-time stochastic linear-quadratic (LQ) control optimal problem with conic control constraints and multiplicative noises on an infinite horizon. There exist in the literature various studies on the estimation and control problems of systems with a multiplicative noise [1,2]. As for the LQ type of stochastic optimal control problems with multiplicative noise, investigations have been focused on the LQ formulation with indefinite penalty matrices on control and state variables for both continuous-time and discrete-time models (see, e.g., [3,4]).

In an LQ optimal problem, the system dynamics are both linear in state and control variables, and the cost function is quadratic in these two variables [5]. One important attractive quality of the LQ type of optimal control models is its explicit control policy which can be derived by solving the corresponding Riccati equation. Due to the elegant structure, the LQ problem has always been a hot issue in optimal control research. Since the fundamental research on deterministic LQ problems by Kalman [6], there have been a great number of studies on it; see [5,7,8]. In the past few years, stochastic LQ problems have drawn more and more attention on this topic, due to the promising applications in different fields, including dynamic portfolio management, financial derivative pricing, population models, and nuclear heat transfer problems; see [9–11].

This paper is motivated by two recent developments: LQ optimal control and Markov decision problems (MDPs). First, the constrained LQ problem is significant in both theory and applications.

Due to the constraints on state and control variables, it is hard to obtain the explicit control policy by solving the Riccati equation [5]. Recently, there have been studies regarding the constrained LQ optimal control problems, such as [12–14]. Meanwhile, in real applications, considering some practical limits, such as the risk or the economic regulations, we have to take some constraints on the control variables into the consideration. In the LQ control problems, including the positivity constraint for the control, some literature, [15,16], propose the optimality conditions and some numerical methods to characterize the optimal control policy. In this paper, we characterise the limits as the conic control constraints considering the real applications.

Work by Cao [17] and Puterman [18] demonstrate that stochastic control problems can be viewed as Markov decision problems. Therefore, the constrained stochastic LQ control problem can be formulated as an MDP, such as [19]. A direct-comparison based approach (or relative optimization), which originated in the area of discrete event systems, has been developed in the past years for the optimization of MDPs [17].

With this approach, optimization is based on the comparison of the performance measures of the system under any two policies. It is intuitively clear, and it can provide new insights, leading to new results to many problems, such as [20–26]. This approach is very convenient and suitable to the performance optimization problems, leading to results including the property of under-selectivity in time-nonhomogeneous Markov processes [24]. In this paper, we show that the special features of the constrained stochastic LQ optimal control make it possible to be solved by the direct-comparison based approach, leading to some new insights for the problem.

In our work, we consider the stochastic LQ control problem through an MDP formulation in the infinite horizon. Through the direct-comparison based approach [17], we first derive the performance potentials for the LQ problem by utilizing the state separation property of the system structure. Based on this, we successfully derive the optimality conditions and the stationary optimal feedback control. We show that the optimal control policy is a piece-wise affine function with respect to the state variables. In real applications, the proposed methodology can be used in many fields, such as system risk contagion [26] and power grid systems [27].

Our work provides a new perspective for LQ control problems. Compared with the former literature, such as [5,13], we still consider the multiplicative noises. We establish a general framework for studying infinite horizon stochastic control problems. With the direct-comparison based approach, which is applicable to both linear and nonlinear systems, we propose more results for the performance optimization problems, and the results can be extended easily. In addition, without identifying all the system parameters, this approach can be implemented on-line, and learning based algorithms can be developed.

The paper is organized as follows. Section 2 introduces an MDP formulation of the constrained stochastic LQ problem with multiplicative noises; some preliminary knowledge on MDP and the state separation property is also provided. In Section 3, we derive the performance difference formula, which is the foundation of the performance optimization; based on it, the Poisson equation and Dynkin's formula can be obtained. Then we derive the optimality condition and the optimal policy through the performance difference formula. In Section 4, we illustrate the results by numerical examples. Finally, we conclude the paper in Section 5.

2. Problem Formulation

2.1. Problem with Infinite Time Horizon

In this section, we study the infinite horizon discrete-time stochastic LQ optimal control problem, in which the conic control constraints are also considered; see [5,14]. For simplicity of the parameters, we consider a one dimensional dynamic system with a multiplicative noise described by

$$x_{l+1} = Ax_l + \mathbf{B}\mathbf{u}_l(x_l) + [Ax_l + \mathbf{B}\mathbf{u}_l(x_l)]\xi_l, \tag{1}$$

for time $l = 0, 1, \cdots$. By denoting $\mathbb{R}$ ($\mathbb{R}_+$) as the set of real (nonnegative real) numbers, in this system, $A \in \mathbb{R}$ and $\mathbf{B} \in \mathbb{R}^{1\times m}$ are deterministic values; $x_l \in \mathbb{R}$ is the state with x_0 being given; and $\mathbf{u}_l \in \mathbb{R}^m$ is a feedback control law at time l. For each l, ξ_l denotes an independent identically distributed one-dimensional multiplicative noise, satisfying a normal distribution with mean 0 and variance $\sigma^2, \sigma \geq 0$. For each l, ξ_l denotes a one-dimensional noise. ξ_l and ξ_k are independent for every $l \neq k$.

Now, we consider the conic control constraint sets (cf. [5])

$$\mathcal{C}_l := \{\mathbf{u}_l | \mathbf{u}_l \in \mathcal{F}_l,\ \mathbf{H}\mathbf{u}_l \in \mathbb{R}^n_+\}, \tag{2}$$

for $l = 0, 1, \cdots$, where $\mathbf{H} \in \mathbb{R}^{n\times m}$ is a deterministic matrix; and $\mathcal{F}_l$ is the filtration of the information available at time l. Let $\mathcal{C}_l \subset R^m$ be a given closed cone; i.e., $\alpha\mathbf{u}_l \in \mathcal{C}_l$ whenever $\mathbf{u}_l \in \mathcal{C}_l$ and $\alpha \geq 0$; and $\mathbf{u}_l + \mathbf{v}_l \in \mathcal{C}_l$ whenever $\mathbf{u}_l, \mathbf{v}_l \in \mathcal{C}$.

The goal of optimization is to minimize the total reward performance measure in a quadratic form:

$$\begin{array}{c}(\mathcal{P}_A)\ \min\limits_{\{\mathbf{u}_l\}|_{l=0}^{\infty}} \eta^{\mathbf{u}} = \lim\limits_{L\to\infty} \mathrm{E}[\sum\limits_{l=0}^{L-1}(Qx_l^2 + \mathbf{u}_l'\mathbf{R}\mathbf{u}_l)|x_0] \\ \text{(s.t.)}\ \{x_t, \mathbf{u}_t\} \text{ satisfies (1) and (2) for } l = 0, 1, \cdots,\end{array} \tag{3}$$

where $Q \in \mathbb{R}_+$ and $\mathbf{R} \in \mathbb{R}^{m\times m}_+$ are deterministic. Here we denote the transpose operation by a prime in the superscript , such as $\mathbf{u}_l'$. $\{\mathbf{u}_l\}$ denotes the control sequence $\{\mathbf{u}_0, \mathbf{u}_1, \cdots\}$. We also assume that (3) exists.

Therefore, the performance function of (3) is

$$f^{\mathbf{u}}(x) = Qx^2 + \mathbf{u}'\mathbf{R}\mathbf{u}. \tag{4}$$

In this paper, we will show that the direct-comparison based approach leads to more new results for the total rewards problem [7], and that the results can be easily extended.

2.2. MDPs with Continuous State Spaces

For a stationary control law $\mathbf{u}_l = \mathbf{u}(x)$, at time $l = 0, 1, \cdots$, the constraint (2) can be written as

$$\mathcal{C} := \{\mathbf{u} | \mathbf{u} \in \mathbb{R}^m,\ \mathbf{H}\mathbf{u} \in \mathbb{R}^n_+\}.$$

Then the above stochastic control problem can be viewed as an MDP with continuous state spaces. More precisely, $\mathbf{u}(x)$ plays a similar role of actions in MDPs, and then the control law $\mathbf{u}$ is the same as a policy.

Consider a discrete-time Markov chain $\mathbf{X} := \{x_l\}_{l=0}^{\infty}$ with a continuous state space on $\mathbb{R}$. The transition probability can be described by a *transition operator* P as

$$(Ph)(x) := \int_{\mathbb{R}} h(y)P(dy|x), \tag{5}$$

where $P(dy|x)$ is the transition probability function, with $x, y \in \mathbb{R}$; and $h(y)$ is any measurable function on $\mathbb{R}$. As ξ_l is independent Gaussian noises, given the current state $x_l = x$, under the stationary control $\mathbf{u}(x), y = x_{l+1}$ satisfies a normal distribution with mean $\mu_y = Ax + \mathbf{B}\mathbf{u}(x)$ and variance $\sigma_y^2 = [Ax + \mathbf{B}\mathbf{u}(x)]^2\sigma^2$. Then we have the transition function of this system as follows,

$$P^{\mathbf{u}}(dy|x) = \frac{1}{\sqrt{2\pi}\sigma_y} exp\{-\frac{(y-\mu_y)^2}{2\sigma_y^2}\}dy. \tag{6}$$

Let $\mathbb{B}$ be the σ-field of $\mathbb{R}$ containing all the (Lebesgue) measurable sets. For any set $B \in \mathbb{B}$, we can define the identity transition function $I(B|x)$. $I(B|x) = 1$ if $x \in B$; $I(B|x) = 0$ otherwise. For any function h and $x \in \mathbb{R}$, we have $(Ih)(x) = h(x)$.

The product of two transition functions $P_1(B|x)$ and $P_2(B|x)$ is defined as a transition function $(P_1P_2)(B|x)$:

$$(P_1P_2)(B|x) := \int_{\mathbb{R}} P_2(B|y)P_1(dy|x),$$

where $x, y \in \mathbb{R}, B \in \mathbb{B}$.

For any transition function P, we can define the kth power, $k = 0, 1, \cdots$, as $P^0 = I, P^1 = P$, and $P^k = PP^{k-1}, k = 2, \cdots$. Suppose that the Markov chain $\mathbf{X}$ is time-homogeneous with transition function $P(B|x), x \in \mathbb{R}, B \in \mathbb{B}$. Then the k-step transition probability functions, denoted as $P^{(k)}(B|x), k = 1, 2, \cdots$, are given by the 1-step transition function defined as $P^{(1)}(B|x) = P(B|x)$ and

$$P^{(k)}(B|x) := \int_{\mathbb{R}} P(dy|x)P^{k-1}(B|y),\ k \geq 2.$$

For any function $h(x)$, we have

$$(P^{(k)}h)(x) = \int_{\mathbb{R}} h(y)P^{(k)}(dy|x) = P(P^{(k-1)})h(x).$$

That is, as an operator, we have $P^{(k)} = P(P^{(k-1)})$. Recursively, we can prove that $P^{(k)} = P^k$.

Suppose that a Markov chain $\mathbf{X}$ with a continuous state space on $\mathbb{R}$ has a steady-state distribution π satisfying $\pi = \pi P$. Define function $e(x) = 1$ for all $x \in \mathbb{R}$. We denote the performance potential g as a function which satisfies the Poisson equation (cf. [17])

$$(I - P)g(x) + \eta(x) = f(x), \tag{7}$$

where I and P are two transition functions, and $\eta(x) = (\pi f)e(x) = \eta e(x)$. Then if g is a solution to (7), so is $g + ce$, with any constant c. We define

$$g_K := \{I + \sum_{k=1}^{K}(P^k - e\pi)\}f, \tag{8}$$

and assume the limit $g(x) := \lim_{K\to\infty} g_K(x)$ exists for $x \in \mathbb{R}$. Then we have the following lemma,

Lemma 1 (Solution to Poisson Equations [17])**.** *For any transition function P and performance function $f(x)$, if*

$$\lim_{k\to\infty} P^k f = (e\pi)f = \eta e,$$
$$\lim_{K\to\infty} g_K = g, \text{ and } \lim_{K\to\infty} Pg_K = Pg,$$

hold for every $x \in \mathbb{R}$, then

$$g = \{I + \sum_{k=1}^{\infty}(P^k - e\pi)\}f, \tag{9}$$

is a solution to the Poisson Equation (7).

2.3. State Separation Property

In order to derive the explicit solution of the stochastic LQ control problem with conic constraints, Reference [14] gives the following lemma for the state separation property of the LQ problem,

Lemma 2 (State Separation [14])**.** *In the system* (1)*, for any $x \in \mathbb{R}$, the optimal solution for problem* (3) *at time l is a piecewise linear feedback policy*

$$\mathbf{u}^*(x_l) = \begin{cases} \hat{\mathbf{K}}^* x_l, & \text{if } x_l \geq 0, \\ -\bar{\mathbf{K}}^* x_l, & \text{if } x_l < 0, \end{cases} \tag{10}$$

for $l = 0, 1, \cdots$, where $\mathcal{K} := \{\mathbf{K} \in \mathbb{R}^m | \mathbf{HK} \in \mathbb{R}^n_+\}$ associated with the control constraint sets $\mathcal{C}_l$; $\hat{\mathbf{K}}^*, \bar{\mathbf{K}}^* \in \mathcal{K}$, are the optimal values of two correspondent auxiliary optimization problems, and the superscript "$*$" corresponds to the optimal control.

Based on (10) in Lemma 2, the stationary control can be written as $\mathbf{u}(x) = \hat{\mathbf{K}} x \mathbf{1}_{x \geq 0} - \bar{\mathbf{K}} x \mathbf{1}_{x<0}$, where $\mathbf{1}_B$ is an indicator function, such that $\mathbf{1}_B = 1$, if the condition B holds true and $\mathbf{1}_B = 0$ otherwise; and $\hat{\mathbf{K}}, \bar{\mathbf{K}} \in \mathcal{K}$. Applying this control, the system dynamics (1) becomes

$$\begin{aligned} x_{l+1} = \ & \hat{C} x_l \mathbf{1}_{x_l \geq 0} + \bar{C} x_l \mathbf{1}_{x_l < 0} \\ & + [\hat{C} x_l \mathbf{1}_{x_l \geq 0} + \bar{C} x_l \mathbf{1}_{x_l < 0}] \xi_l, \end{aligned} \tag{11}$$

for $l = 0, 1, \cdots$, where

$$\hat{C} = A + \mathbf{B}\hat{\mathbf{K}},\ \bar{C} = A - \mathbf{B}\bar{\mathbf{K}}. \tag{12}$$

Moreover, the performance measure (3) becomes

$$\eta^{\mathbf{u}}(x) = \lim_{L \to \infty} \mathrm{E}[\sum_{l=0}^{L-1} \hat{W} x_l^2 \mathbf{1}_{x_l \geq 0} + \bar{W} x_l^2 \mathbf{1}_{x_l < 0} | x_0 = x],$$

where $\hat{W} = Q + \hat{\mathbf{K}}^T \mathbf{R} \hat{\mathbf{K}}$ and $\bar{W} = Q + \bar{\mathbf{K}}^T \mathbf{R} \bar{\mathbf{K}}$. Therefore, the performance function (4) becomes

$$f(x) = \hat{W} x^2 \mathbf{1}_{x \geq 0} + \bar{W} x^2 \mathbf{1}_{x < 0}. \tag{13}$$

It is easy to verify that $\hat{W}$ and $\bar{W}$ are positive semi-definite. We assume that this one-dimensional state system is stable, and then the spectral radiuses of $\hat{C}$ and $\bar{C}$ are less than 1, i.e., $C^{max} = max(\hat{C}, \bar{C}) < 1$. In the next section, we will derive the performance potentials for the LQ problem, which is the foundation of the performance optimization. Based on this, the Poisson equation and the Dynkin's formula can be derived. The direct-comparison based approach provides a new perspective for this problem,and the results can be extended easily.

3. Performance Optimization

In this section, utilizing the state separation property, we derive the performance difference formula, which compares the performance measures of any tow policies, and then derive the optimality condition and the optimal policy with the direct-comparison based approach.

3.1. Performance Difference Formula

We denote $\hat{W}_0 = \hat{W}$ and $\bar{W}_0 = \bar{W}$. Then we have the performance function as $f(x) = \hat{W}_0 x^2 \mathbf{1}_{x\geq 0} + \bar{W}_0 x^2 \mathbf{1}_{x<0}$. With the initial condition $x_0 = x$, by (5), (6), (11), and (13), the performance operator is

$$(Pf)(x) = \hat{W}_1 x^2 \mathbf{1}_{x\geq 0} + \bar{W}_1 x^2 \mathbf{1}_{x<0}, \tag{14}$$

where

$$\hat{W}_1 = (a_1\hat{W}_0 + a_2\bar{W}_0)\hat{C}^2,\ \bar{W}_1 = (a_1\hat{W}_0 + a_2\bar{W}_0)\bar{C}^2,$$

and

$$a_1 = \sigma\phi(\frac{1}{\sigma}) + (1+\sigma^2)\Phi(\frac{1}{\sigma}), \tag{15}$$

$$a_2 = -\sigma\phi(-\frac{1}{\sigma}) + (1+\sigma^2)\Phi(-\frac{1}{\sigma}), \tag{16}$$

with $\phi(\cdot)$ as the probability density function of a standard normal distribution. We can verify that a_1 and a_2 are both nonnegative constants, with $a_1 + a_2 = 1 + \sigma^2$.

As $P^2 f = P(Pf)$, continuing this process, we obtain

$$(P^k f)(x) = \hat{W}_k x^2 \mathbf{1}_{x\geq 0} + \bar{W}_k x^2 \mathbf{1}_{x<0}, \tag{17}$$

where

$$\hat{W}_k = (a_1\hat{W}_{k-1} + a_2\bar{W}_{k-1})\hat{C}^2,$$
$$\bar{W}_k = (a_1\hat{W}_{k-1} + a_2\bar{W}_{k-1})\bar{C}^2.$$

We set $W_0^* = \max(\hat{W}_0, \bar{W}_0)$.

In order to ensure the stability of the system, Reference [14] gives some assumptions. Here we assume $\max(\hat{C}^2, \bar{C}^2) < 1/(1+\sigma^2) \leq 1$. Then we have

$$\hat{W}_k \leq (1+\sigma^2)^k (\hat{C}^2)^k W_0^*,\ \bar{W}_k \leq (1+\sigma^2)^k (\bar{C}^2)^k W_0^*.$$

Therefore, we have

$$\lim_{k\to+\infty} \hat{W}_k = \lim_{k\to+\infty} \bar{W}_k = 0. \tag{18}$$

We denote $\hat{\mathbf{G}}_k := \sum_{i=0}^k \hat{W}_i$ and $\bar{\mathbf{G}}_k := \sum_{i=0}^k \bar{W}_i$. Based on the above claims, we obtain that $\hat{\mathbf{G}}_k$ and $\bar{\mathbf{G}}_k$ would converge when $k \to +\infty$. Thus we denote

$$\hat{\mathbf{G}} := \lim_{K\to+\infty} \hat{\mathbf{G}}_K = \sum_{k=0}^{+\infty} \hat{W}_k,\ \bar{\mathbf{G}} := \lim_{K\to+\infty} \bar{\mathbf{G}}_K = \sum_{k=0}^{+\infty} \bar{W}_k.$$

Based on the definition of total rewards (3), we have

$$\eta(x) = \hat{\mathbf{G}} x^2 \mathbf{1}_{x\geq 0} + \bar{\mathbf{G}} x^2 \mathbf{1}_{x<0}. \tag{19}$$

By (17) and (18), we have

$$\lim_{k\to+\infty} (P^k f)(x) = 0.$$

Then we have proved that the closed-loop system (11) is L^2-asymptotically stable, i.e., $\lim_{l\to\infty} E[(x_l)^2] = 0$. Therefore, the total rewards $\eta(x)$ exists, that is, a piecewise quadratic function with positive semi-definite matrices $\hat{\mathbf{G}}$ and $\bar{\mathbf{G}}$.

Now, we define the discrete version of generator, $\mathcal{A}$ for any function $h(x), x \in \mathbb{R}$, such that

$$\mathcal{A}h(x) := (Ph)(x) - h(x). \tag{20}$$

Taking $h(x)$ as $\eta(x)$, and by the definition of $\eta(x)$ in (3), we have the *Poisson equation* as follows,

$$\mathcal{A}\eta(x) + f(x) = 0. \tag{21}$$

By (5) and (20), we obtain the discrete version of *Dynkin's formula* as

$$E\{\sum_{k=0}^{K-1}[\mathcal{A}h(x_k)]|x_0\} = E\{h(x_K)|x_0\} - h(x_0). \tag{22}$$

and if the limit $K \to \infty$ exists, then

$$E\{\sum_{k=0}^{\infty}[\mathcal{A}h(x_k)]|x_0\} = \lim_{K\to\infty} E\{h(x_K)|x_0\} - h(x_0).$$

Now, we consider two policies $\mathbf{u}, \mathbf{u}' \in \mathcal{U}_0$, resulting in two independent Markov chains $\mathbf{X}$ and $\mathbf{X}'$ in the same state space $\mathbb{R}$, with $P, f, \eta, \mathcal{A}, E$, and $P', f', \eta', \mathcal{A}', E'$, respectively. Let $x'_0 = x_0$. Applying the Dynkin's Formula (22) on $\mathbf{X}'$ with $h(x) = \eta(x)$ yields

$$E'\{\sum_{k=0}^{K-1}[\mathcal{A}'\eta(x'_k)|x_0\} = E'\{\eta(x_K)]|x_0\} - \eta(x_0). \tag{23}$$

Noting that $\eta'(x_0) = \lim_{K\to\infty}\sum_{k=0}^{K-1}\{E'[f'(x_k)]|x_0\}$, and $\lim_{K\to\infty} E'\{\eta(x_K)|x_0\} = 0$ due to asymptotical stability. Then by (23), we obtain the performance difference formula:

$$\eta'(x_0) - \eta(x_0) = \lim_{K\to\infty}\sum_{k=0}^{K-1} E'\{(\mathcal{A}'\eta + f')(x'_k)|x_0\}. \tag{24}$$

3.2. Optimal Policy

Based on the performance difference Formula (24), we have the following optimality condition.

Theorem 1 (Optimality Condition)**.** *A policy* $\mathbf{u}^*$ *in* $\mathcal{C}$ *is optimal if, and only if,*

$$\mathcal{A}^{\mathbf{u}}\eta^{\mathbf{u}^*} + f^{\mathbf{u}} \geq 0 = \mathcal{A}^{\mathbf{u}^*}\eta^{\mathbf{u}^*} + f^{\mathbf{u}^*}, \forall \mathbf{u} \in \mathcal{C}. \tag{25}$$

From (25), the optimality equation is:

$$\min_{\mathbf{u}\in\mathcal{C}}\{\mathcal{A}^{\mathbf{u}}\eta^{\mathbf{u}^*} + f^{\mathbf{u}}\} = 0. \tag{26}$$

Proof. First, the "if" part follows from the performance difference Formula (24) and the Poisson Equation (21).

Next, we prove the "only if" part: Let $\mathbf{u}^*$ be an optimal policy. We need to prove that (25) holds. Suppose that this is not true. Then, there must exist one policy, denoted as $\mathbf{u}'$, such that (25) does not hold. That is, there must be at least one state, denoted as y, such that

$$P^{\mathbf{u}^*}\eta^{\mathbf{u}^*}(y) + f^{\mathbf{u}^*}(y) > P^{\mathbf{u}'}\eta^{\mathbf{u}^*}(y) + f^{\mathbf{u}'}(y).$$

Then we can create a policy $\tilde{\mathbf{u}}$ by setting $\tilde{\mathbf{u}} = \mathbf{u}'$ when $x = y$, and $\tilde{\mathbf{u}} = \mathbf{u}^*$ when $x \neq y$. We have $\eta^{\mathbf{u}^*} > \eta^{\mathbf{u}'}$. This contradicts to the fact that $\mathbf{u}^*$ is an optimal policy. □

Based on the optimality condition, the optimal control $\mathbf{u}^*$ can be obtained by developing policy iteration algorithms. Roughly speaking, we start with any policy $\mathbf{u}_0$. At the kth step, $k = 0, 1, \cdots$, given a piecewise linear policy $\mathbf{u}_k(x) = \hat{\mathbf{K}}x\mathbf{1}_{x\geq 0} - \bar{\mathbf{K}}x\mathbf{1}_{x<0}$, where $\hat{\mathbf{K}}, \bar{\mathbf{K}} \in \mathcal{K}$, we want to find a better policy by (26). We consider any policy $\mathbf{u}(x)$. Setting $h(x) = \eta^{\mathbf{u}_k}(x) = \hat{\mathbf{G}}x^2\mathbf{1}_{x\geq 0} + \bar{\mathbf{G}}x^2\mathbf{1}_{x<0}$, by (5), (12), and (14), we have

$$\begin{aligned}(P^{\mathbf{u}}\eta^{\mathbf{u}_k})(x) =& (a_1\hat{\mathbf{G}} + a_2\bar{\mathbf{G}})(A + \mathbf{B}\hat{\mathbf{K}})^2x^2\mathbf{1}_{x\geq 0} \\ &+ (a_1\hat{\mathbf{G}} + a_2\bar{\mathbf{G}})(A - \mathbf{B}\bar{\mathbf{K}})^2x^2\mathbf{1}_{x<0}.\end{aligned} \tag{27}$$

where a_1 and a_2 satisfy Equations (15) and (16), respectively.

Then, from (4) and (27), we have

$$\begin{aligned}\mathbf{u}_{k+1}(x) &= \arg\{\min_{\mathbf{u}\in\mathcal{C}}[(P^{\mathbf{u}}\eta^{\mathbf{u}_k})(x) + f^{\mathbf{u}}(x)]\} \\ &= \hat{\mathbf{K}}_{k+1}x\mathbf{1}_{x\geq 0} - \bar{\mathbf{K}}_{k+1}x\mathbf{1}_{x<0},\end{aligned}$$

with

$$\begin{aligned}\hat{\mathbf{K}}_{k+1} &= \arg\min_{\mathbf{K}\in\mathcal{K}}[a_1\hat{C}^2\hat{\mathbf{G}} + a_2\hat{C}^2\bar{\mathbf{G}} + Q + \mathbf{K}^T\mathbf{R}\mathbf{K}], \\ \bar{\mathbf{K}}_{k+1} &= \arg\min_{\mathbf{K}\in\mathcal{K}}[a_1\bar{C}^2\hat{\mathbf{G}} + a_2\bar{C}^2\bar{\mathbf{G}} + Q + \mathbf{K}^T\mathbf{R}\mathbf{K}],\end{aligned}$$

where $\hat{C} = A + \mathbf{BK}$, and $\bar{C} = A - \mathbf{BK}$.

It can be seen that if the policy $\mathbf{u}_k(x)$ is a piecewise linear control, then we can find an improved policy $\mathbf{u}_{k+1}(x)$, which is also piecewise linear. Moreover, if $\hat{\mathbf{K}}_{k+1} = \hat{\mathbf{K}}$ and $\bar{\mathbf{K}}_{k+1} = \bar{\mathbf{K}}$, i.e., $\mathbf{u}_{k+1} = \mathbf{u}_k$, then the iteration stops. The policy $\mathbf{u}_k$ satisfies the optimal condition (26) in Theorem 1, and therefore is an optimal control.

Therefore, we can obtain the optimal policy as follows,

$$\mathbf{u}^*(x) = \hat{\mathbf{K}}^*x\mathbf{1}_{x\geq 0} - \bar{\mathbf{K}}^*x\mathbf{1}_{x<0}, \tag{28}$$

where

$$\hat{\mathbf{K}}^* = \arg\min_{\mathbf{K}\in\mathcal{K}}[a_1\hat{C}^2\hat{\mathbf{G}}^* + a_2\hat{C}^2\bar{\mathbf{G}}^* + Q + \mathbf{K}^T\mathbf{R}\mathbf{K}], \tag{29}$$

$$\bar{\mathbf{K}}^* = \arg\min_{\mathbf{K}\in\mathcal{K}}[a_1\bar{C}^2\hat{\mathbf{G}}^* + a_2\bar{C}^2\bar{\mathbf{G}}^* + Q + \mathbf{K}^T\mathbf{R}\mathbf{K}]. \tag{30}$$

Moreover,

$$\hat{\mathbf{G}}^* = \min_{\mathbf{K}\in\mathcal{K}}\{a_1\hat{C}^2\hat{\mathbf{G}}^* + a_2\hat{C}^2\bar{\mathbf{G}}^* + Q + \mathbf{K}^T\mathbf{R}\mathbf{K}\}, \tag{31}$$

$$\bar{\mathbf{G}}^* = \min_{\mathbf{K}\in\mathcal{K}}\{a_1\bar{C}^2\hat{\mathbf{G}}^* + a_2\bar{C}^2\bar{\mathbf{G}}^* + Q + \mathbf{K}^T\mathbf{R}\mathbf{K}\}. \tag{32}$$

The original problem (3) is transformed to two auxiliary optimization problems (29) and (30). Under the optimal control $\mathbf{u}^*$ in (28), the closed-loop system (11) is L^2-asymptotically stable. From (19), with the initial condition $x_0 = x$, we know the optimal total reward performance is

$$\eta^*(x) = \hat{\mathbf{G}}^*x^2\mathbf{1}_{x\geq 0} + \bar{\mathbf{G}}^*x^2\mathbf{1}_{x<0}, \tag{33}$$

where $\hat{\mathbf{G}}^*$ and $\bar{\mathbf{G}}^*$ satisfy (31) and (32), respectively.

Policy iteration can also be implemented on-line, the performance (potential) can be learned on a sample path without knowing all the transition probabilities. In on-line algorithms, the computation of policy evaluation is $O(n)$, where n is the length of a sample path. Additionally, Reference [14] also provides some algorithms for calculating the optimal policy.

4. Simulation Examples

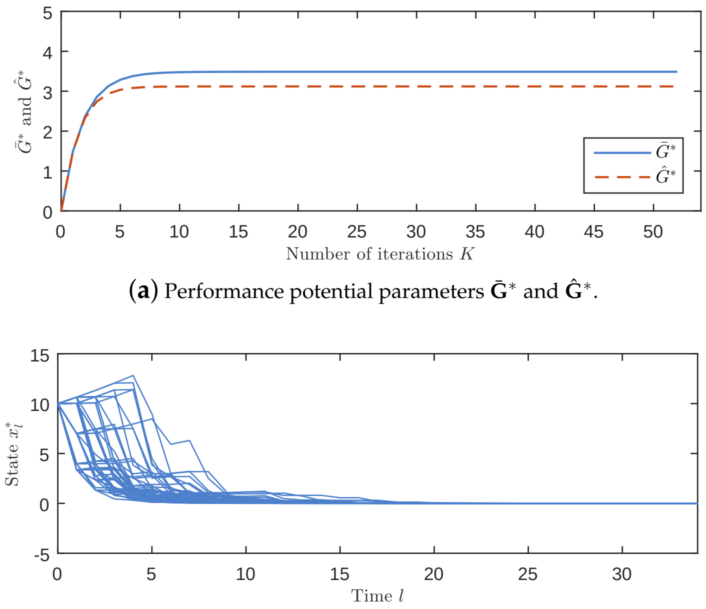
In this section, we use two numerical examples to illustrate the optimal policy for the constrained LQ control problem (3).

Example 1. *We consider a stochastic LQ system with* $x_0 = 10$, $m = 3$, $A = 0.8$, *and* $B = (-0.35, 0.18, 0.25)'$. *The cost matrix is*

$$\mathbf{R} = \begin{pmatrix} 1.2 & 0.6 & 0.4 \\ 0.6 & 1.8 & 0.2 \\ 0.4 & 0.2 & 2.4 \end{pmatrix}, \textit{ and } Q = 1.2.$$

For time $l = 0, 1, \cdots$, *the variance of the 0-mean i.i.d. Gaussian noise* ξ_l *is* $\sigma^2 = 0.25$. *We consider the conic constraint* $\mathbf{u} \geq 0$. *By applying Theorem 1, the stationary optimal control is* $\mathbf{u}_l^*(x_l) = \hat{\mathbf{K}}^* x_l \mathbf{1}_{x_l \geq 0} - \bar{\mathbf{K}}^* x_l \mathbf{1}_{x_l < 0}$, *for* $l = 0, 1, \cdots$, *where* $\hat{\mathbf{K}}^* = (0.574, 0, 0)'$, $\bar{\mathbf{K}}^* = (0, 0.250, 0.270)'$, $\hat{\mathbf{G}}^* = 2.773$ *and* $\bar{\mathbf{G}}^* = 3.473$. *Furthermore, the optimal reward performance is* $\eta^*(x_0) = \hat{\mathbf{G}}^* x_0^2 \mathbf{1}_{x_0 \geq 0} + \bar{\mathbf{G}}^* x_0^2 \mathbf{1}_{x_0 < 0} = 623.987$.

As shown in Figure 1a plots the outputs $\bar{\mathbf{G}}^*$ and $\hat{\mathbf{G}}^*$ with respect to iteration time K; Figure 1b plots the state trajectories of 50 samples by setting $x_0 = 10$ and implementing the stationary optimal control $\mathbf{u}^*$. It can be observed that x_l^* converges to 0 after time $l = 20$ and this closed loop system is asymptotically stable.



(**a**) Performance potential parameters $\bar{\mathbf{G}}^*$ and $\hat{\mathbf{G}}^*$.

(**b**) State trajectories of 50 samples under the optimal control $\mathbf{u}^*$.

Figure 1. The simulation results of Example 1.

Example 2. *In the second case, we assume $x_0 = 10$, A and B, following the identical discrete distribution with five cases. We assume $A \in (-0.7, -0.6, 0.9, 1, 1.1)$, and*

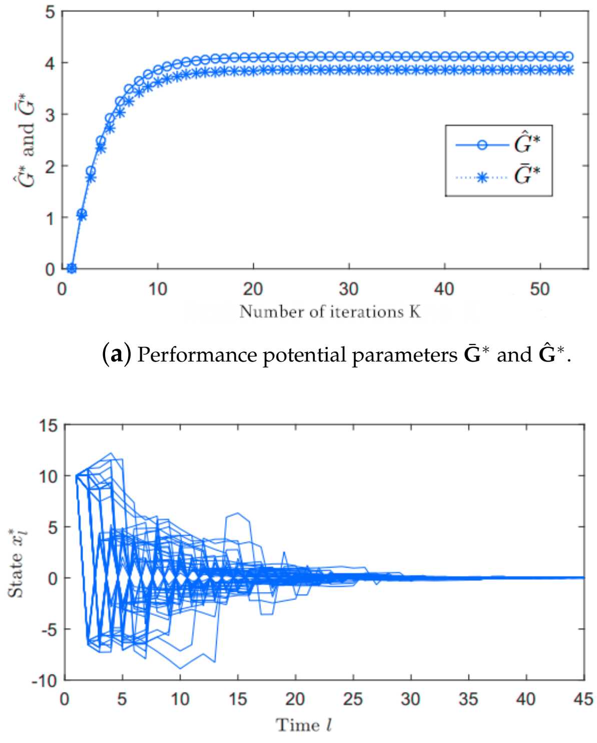
$$B \in \left\{ \begin{pmatrix} 0.18 \\ -0.05 \\ -0.140 \end{pmatrix}, \begin{pmatrix} 0.03 \\ -0.12 \\ -0.03 \end{pmatrix}, \begin{pmatrix} -0.05 \\ 0.05 \\ 0.05 \end{pmatrix}, \begin{pmatrix} -0.01 \\ 0.05 \\ 0.01 \end{pmatrix}, \begin{pmatrix} -0.05 \\ 0.01 \\ 0.06 \end{pmatrix} \right\},$$

each of which has the same probability 0.2. The cost matrix is

$$\mathbf{R} = \begin{pmatrix} 1.5 & 0.6 & 0.4 \\ 0.6 & 1.5 & 0.2 \\ 0.4 & 0.2 & 2.5 \end{pmatrix}, \text{ and } Q = 1.5.$$

For time $l = 0, 1, \cdots$, the variance of the 0-mean i.i.d. Gaussian noise ξ_l is $\sigma^2 = 0.25$. We consider the conic constraint $\mathbf{u} \geq 0$. By applying Theorem 1, the stationary optimal control is $\mathbf{u}_l^*(x_l) = \hat{\mathbf{K}}^* x_l \mathbf{1}_{x_l \geq 0} - \bar{\mathbf{K}}^* x_l \mathbf{1}_{x_l < 0}$, for $l = 0, 1, \cdots$, where $\hat{\mathbf{K}}^*$ and $\bar{\mathbf{K}}^*$ are identified as follows, $\hat{\mathbf{K}}^* = (0.259, 0.100, 0.130)'$, $\bar{\mathbf{K}}^* = (0.100, 0.500, 0.100)'$, $\hat{\mathbf{G}}^* = 4.111$ and $\bar{\mathbf{G}}^* = 3.859$. Furthermore, the optimal reward performance is $\eta^*(x_0) = \hat{\mathbf{G}}^* x_0^2 \mathbf{1}_{x_0 \geq 0} + \bar{\mathbf{G}}^* x_0^2 \mathbf{1}_{x_0 < 0} = 489.23$.

As shown in Figure 2a plots the outputs $\bar{\mathbf{G}}^*$ and $\hat{\mathbf{G}}^*$ with respect to iteration times K; Figure 2b plots the state trajectories of 50 samples by setting $x_0 = 10$ and implementing the stationary optimal control $\mathbf{u}^*$. It can be observed that x_l^* converges to 0 after time $l = 35$ and this closed loop system is asymptotically stable.



(**a**) Performance potential parameters $\bar{\mathbf{G}}^*$ and $\hat{\mathbf{G}}^*$.

(**b**) State trajectories of 50 samples under the optimal control $\mathbf{u}^*$.

Figure 2. Simulation Results of Example 2.

5. Conclusions

In this paper, we apply the direct-comparison based optimization approach to study the rewards optimization of the discrete-time stochastic linear-quadratic control problem with conic constraints on an infinite horizon. We derive the performance difference formula by utilizing the state separation property of the system structure. Based on this, the optimality condition and the stationary optimal feedback control can be obtained. The direct-comparison based approach is applicable to both linear and nonlinear systems. By introducing the the LQ optimization problem, we establish a general framework for studying infinite horizon control problems with total rewards. We verify that the proposed optimal approach can solve the LQ problems. Then we illustrate our results by two simulation examples.

The results can easily be extended to the cases of non-Gaussian noises and average rewards. Most significantly, our methodology can deal with a very general class of linear constraints on state and control variables, which includes the cone constraints, positivity and negativity constraints, and the state-dependent upper and lower bound constraints as a special case. In addition to the problem with the infinite control horizon, our results still fit problems with a finite horizon. In addition, without identifying all the system structure parameters, this approach can also be implemented on-line, and learning based algorithms can be developed.

Finally, this work focuses on the discrete-time stochastic LQ control problem. Our next step is to investigate continuous cases. As the constrained LQ problem has a wide range of applications, we hope to apply our approach in more areas, such as dynamic portfolio management, security optimization of cyber-physical systems, and financial derivative pricing, in our future research.

Author Contributions: Conceptualization, R.X. and X.Y.; methodology, R.X.; validation, R.X., X.Y. and W.W.; formal analysis, X.Y.; data curation, X.Y.; writing–original draft preparation, R.X. and X.Y.; writing–review and editing, R.X., X.Y. and W.W.; supervision, W.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under Grant 61573244 "Stochastic control optimization of uncertain systems based on offset with multiplicative noises and its applications in the financial optimization" and 61521063 "Control theory and techniques: design, control and optimization of network systems".

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MDP	Markov Decision Process
LQ	Linear-Quadratic

References

1. Basin, M.; Perez, J.; Skliar, M. Opitmal filtering for polynomial system wtates with polynomial multiplicative noise. *Int. J. Robust Nonlinear Control* **2006**, *16*, 303–314. [CrossRef]
2. Gershon, E.; Shaked, U. Static H2 and Houtput-feedback of discrete-time LTI systems with state multiplicative noise. *Syst. Control Lett.* **2006**, *55*, 232–239. [CrossRef]
3. Lim, A.B.E.; Zhou, X.Y. Stochastic optimal control LQR control with integral quadratic constraints and indefinite control weights. *IEEE Trans. Autom. Control* **1999**, *44*, 1359–1369. [CrossRef]
4. Zhu, J. On stochastic riccati equations for the stochastic LQR problem. *Syst. Control Lett.* **2005**, *44*, 119–124. [CrossRef]
5. Hu, Y.; Zhou, X.Y. Constrained stochastic LQ control with random coefficients, and application to portfolio selection. *SIAM J. Control Optim.* **2005**, *44*, 444–446. [CrossRef]
6. Kalman, R.E. Contributions to the theory of optimal control. *Bol. Soc. Mat. Mex.* **1960**, *5*, 102–119.
7. Anderson, B.D.; Moore, J.B. *Optimal Control: Linear Quadratic Methods*; Courier Corporation: North Chelmsford, MA, USA, 2007; pp. 167–189.

8. Yong, J. Linear-quadratic optimal control problems for mean-field stochastic differential equations. *SIAM J. Control Optim.* **2013**, *51*, 2809–2838. [CrossRef]
9. Gao, J.J.; Li, D.; Cui, X.Y.; Wang, S.Y. Time cardinality constrained mean-variance dynamic portfolio selection and market timing: A stochastic control approach. *Automatica* **2015**, *54*, 91–99. [CrossRef]
10. Costa, O.L.V.; Fragoso, M.D.; Margues, R.P. *Discrete-Time Markov Jump Linear Systems*; Springer: Berlin, Germany, 2007; pp. 291–317.
11. Primbs, J.A.; Sung, C.H. Stochastic receding horizon control of contrained linear systems with state and control multiplicative noise. *IEEE Trans. Autom. Control* **2009**, *54*, 221–230. [CrossRef]
12. Dong, Y.C. Constrained LQ problem with a random jump and application to portfolio selection . *Chin. Ann. Math.* **2019**, *39*, 829–848. [CrossRef]
13. Gao, J.J.; Li, D. Cardinality constrained linear quadratic optimal control. *IEEE Trans. Autom. Control* **2011**, *56*, 1936–1941. [CrossRef]
14. Wu, W.P.; Gao, J.J.; Li, D.; Shi, Y. Explicit solution for constrained stochastic linear-quadratic control with multiplicative noise. *IEEE Trans. Autom. Control* **2019**, *64*, 1999–2012. [CrossRef]
15. Campbell, S.L. On positive controllers and linear quadratic optimal control problems. *Int. J. Control* **1982**, *36*, 885–888. [CrossRef]
16. Heemels, W.P.; Eijndhoven, S.V.; Stoorvogel, A.A. Linear quadratic regulator problem with positive controls. *Int. J. Control* **1998**, *70*, 551–578. [CrossRef]
17. Cao, X.R. *Stochastic Learning and Optimization: A Sensitivity-Based Approach*; Springer: New York, NY, USA, 2007.
18. Puterman, M.L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*; Wiley: New York, NY, USA, 1994.
19. Chen, R.C. Constrained stochastic control with probabilistic criteria and search optimization. In Proceedings of the 43rd IEEE Conference on Decision and Control (CDC), Nassau, Bahamas, 14–17 December 2004.
20. Zhang, K.J.; Xu, Y.K.; Chen, X.; Cao, X.R. Policy iteration based feedback control. *Automatica* **2008**, *44*, 1055–1061. [CrossRef]
21. Cao, X.R. Stochastic feedback control with one-dimensional degenerate diffusions and nonsmooth value functions. *IEEE Trans. Autom. Control* **2018**, *62*, 6136–6151. [CrossRef]
22. Cao, X.R.; Wan, X.W. Sensitivity analysis of nonlinear behavior with distorted probability. *Math. Financ.* **2017**, *27*, 115–150. [CrossRef]
23. Xia, L. Mean-variance optimization of discrete time discounted Markov decision processes. *Automatica* **2018**, *88*, 76–82. [CrossRef]
24. Cao, X.R. Optimality consitions for long-run average rewards with underselectivity and nonsmooth features. *IEEE Trans. Autom. Control* **2017**, *62*, 4318–4332. [CrossRef]
25. Xue, R.B.; Ye, X.S.; Cao, X.R. Optimization of stock trading with additional information by Limit Order Book. *Automatica* **2019**, submitted.
26. Ye, X.S.; Xue, R.B.; Gao, J.J.; Cao, X.R. Optimization in curbing risk contagion among financial institutes. *Automatica* **2018**, *94*, 214–220. [CrossRef]
27. Jia, Q.S.; Yang, Y.; Xia, L.; Guan, X.H. A tutorial on event-based optimization with application in energy Internet. *J. Control Theory Appl.* **2018**, *35*, 32–40.