Total Rewards Optimization of Conic Constrained Stochastic Linear-Quadratic Control: A Direct-Comparison Based Approach

Xiang-Shen Ye[1], Ruo-Bing Xue[1], and Weiping Wu[1]

***Abstract*— In this paper, we study the optimization of the discrete-time stochastic linear-quadratic (LQ) control problem with conic control constraints in infinite horizon. 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 the total rewards optimization problem. By utilizing the state separation property, we successfully derive the performance difference formula. Based on it, the optimality condition and the stationary optimal feedback control can be obtained. We show that the optimal control policy is a piece-wise affine function with respect to the state variables. Our work provides a new perspective in LQ control problems. The direct-comparison based approach is applicable to both linear and nonlinear systems. Based on this approach, learning based algorithms can be developed without identifying all the system parameters.**

I. Introduction

In this paper, we study the total rewards optimization of the discrete-time stochastic linear-quadratic (LQ) control problem with conic control constraints in infinite horizon. In an LQ optimal problem, the system dynamics are both linear in state and control variables, and the cost functional is quadratic in these two variables [1]. Because of 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 [2], there have been a great number of researches on it; see [1], [3], and [4].

This paper is motivated by two recent developments in LQ optimal control and Markov decision problems (MDPs). First, the constrained LQ problem is significant in both theory and applications, such as the no shorting constraint in portfolio, and upper/lower bounds for control variables. However, because of the constraints on state and control variables, it is hard to obtain the explicit control policy by solving the Riccati equation [1]. Recently, there are some promising researches about constrained LQ optimal control problems, such as [1], [5], and [6].

On the other hand, as we know, stochastic control problems can be viewed as Markov decision problems; see [7] and [8]. Therefore, the constrained stochastic LQ control problem can be formulated as an MDP. On this side, a direct-comparison based approach has been developed in the past years to the optimization of MDPs [7]. 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 [9], [10], [11], [12], and [13]. 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 to the problem.

In this paper, we consider the total reward MDP problem in infinite horizon. Through the direct-comparison based approach, we first derive the Poisson equation and the Dynkin's formula by utilizing the state separation property of the system structure. Then we successfully derive the performance difference formula. Based on it, the optimality condition and the stationary optimal feedback control can be obtained. We show that the optimal control policy is a piece-wise affine function with respect to the state variables. Our work provides a new perspective in LQ control problems. The direct-comparison based approach is applicable to both linear and nonlinear systems. 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 II introduces an MDP formulation of the constrained stochastic LQ problem; some preliminary knowledge on MDP and the state separation property is also provided. In Section III, we derive the performance difference formula; based on it, the optimality condition and the optimal policy can be obtained. In Section IV, we illustrate the results by a numerical example. Finally, we conclude the paper in Section V.

II. Problem Formulation

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 [1], [6]. For simplicity of parameters, we consider a one dimensional dynamic system with a multiplicative noise described by

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

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 $\sigma^2, \sigma \geq 0$.

Now we consider the conic control constraint sets (cf. [1])

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

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.

[1]The authors are with the Department of Automation, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, China. Emails: {yesyjs, cynthiabaobei, godream}@sjtu.edu.cn

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:

$$\min_{\{\mathbf{u}_l\}} \eta^{\{\mathbf{u}_l\}}(x) = \lim_{L\to\infty} \mathrm{E}[\sum_{l=0}^{L-1} (Qx_l^2 + \mathbf{u}_l^T \mathbf{R} \mathbf{u}_l) | x_0 = x] \quad (3)$$

$$\text{(s.t.)} \quad \{x_l, \mathbf{u}_l\} \text{ satisfies (1) and (2) for } l = 0, 1, \cdots,$$

where $Q \in \mathbb{R}_+$ and $\mathbf{R} \in \mathbb{R}_+^{m\times m}$ are deterministic. Here we denote the transpose operation by a superscript "T", such as $\mathbf{u}_l^T$. And $\{\mathbf{u}_l\}$ denotes the control sequence $\{\mathbf{u}_0, \mathbf{u}_1, \cdots\}$.

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

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

Therefore, the performance function of (3) is

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

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

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

where $P(dy|x)$ is the transition probability function, with $x, y \in \mathbb{R}$; and $h(y)$ is any measurable function on $\mathbb{R}$.

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 $P^{(k)} = P^k$.

Because ξ_l is an independent Gaussian noise, 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,

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

In order to derive the explicit solution of the constrained stochastic LQ control problem, [6] gives the following lemma for the state separation property of the LQ problem,

Lemma 1 (State Separation [6]): For any $x \in \mathbb{R}$, the optimal solution for problem (3) at time l is a piecewise linear feedback policy

$$\mathbf{u}^*(x_l) = \begin{cases} \hat{\mathbf{K}}^* x_l, & \text{if } x_l \geq 0, \\ -\bar{\mathbf{K}}^* x_l, & \text{if } x_l < 0, \end{cases} \quad (7)$$

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

Based on (7) in Lemma 1, 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

$$\begin{aligned} x_{l+1} = \ & \hat{C} x_l \mathbf{1}_{x_l \geq 0} + \bar{C} x_l \mathbf{1}_{x_l < 0} \\ & + [\hat{C} x_l \mathbf{1}_{x_l \geq 0} + \bar{C} x_l \mathbf{1}_{x_l < 0}] \xi_l, \end{aligned} \quad (8)$$

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

$$\hat{C} = A + \mathbf{B}\hat{\mathbf{K}}, \ \bar{C} = A - \mathbf{B}\bar{\mathbf{K}}. \quad (9)$$

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

$$f(x) = \hat{W} x^2 \mathbf{1}_{x\geq 0} + \bar{W} x^2 \mathbf{1}_{x<0}. \quad (10)$$

It is easy to verify that $\hat{W}$ and $\bar{W}$ are positive semi-definite.

With the dynamic programming approach, the total reward performance (3) of the system is discussed in [6]. In this paper, we will show that the direct-comparison based approach provides a new perspective for this problem, and the results can be extended easily. In the next section, we will derive the optimal policy for the LQ problem.

III. Optimization of Total Rewards

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.

A. Performance Difference Formula

Denote $\hat{W}_0 = \hat{W}$ and $\bar{W}_0 = \bar{W}$. Then we have the performance function as $f(x) = \hat{W}_0 x^2 \mathbf{1}_{x\geq 0} + \bar{W}_0 x^2 \mathbf{1}_{x<0}$. From (5), (6), (8), and (10), the performance operator is

$$(Pf)(x) = \hat{W}_1 x^2 \mathbf{1}_{x\geq 0} + \bar{W}_1 x^2 \mathbf{1}_{x<0}, \tag{11}$$

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

$$a_1 = \sigma\phi(\frac{1}{\sigma}) + (1+\sigma^2)\Phi(\frac{1}{\sigma}), \tag{12}$$

$$a_2 = -\sigma\phi(-\frac{1}{\sigma}) + (1+\sigma^2)\Phi(-\frac{1}{\sigma}), \tag{13}$$

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 $a_1 + a_2 = 1 + \sigma^2$.

Continuing this process, we obtain

$$(P^k f)(x) = \hat{W}_k x^2 \mathbf{1}_{x\geq 0} + \bar{W}_k x^2 \mathbf{1}_{x<0}, \tag{14}$$

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, [6] 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

$$\lim_{k\to+\infty} \hat{W}_k = \lim_{k\to+\infty} \bar{W}_k = 0. \tag{15}$$

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

$$\eta(x) = \hat{\mathbf{G}}x^2\mathbf{1}_{x\geq 0} + \bar{\mathbf{G}}x^2\mathbf{1}_{x<0}. \tag{16}$$

By (14) and (15), we have

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

Then we have proved that the closed-loop system (8) 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

$$\mathcal{A}h(x) := (Ph)(x) - h(x). \tag{17}$$

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

$$\mathcal{A}\eta(x) + f(x) = 0. \tag{18}$$

By (5) and (17), we obtain the *Dynkin's formula* as

$$E\{\sum_{k=0}^{K-1} [\mathcal{A}h(x_k)]|x_0\} = E\{h(x_K)|x_0\} - h(x_0). \tag{19}$$

Now we consider two policies $\mathbf{u}, \mathbf{u}' \in \mathcal{U}_0$, with two Markov chains 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 (19) on $\mathbf{X}'$ with $h(x) = \eta(x)$ yields

$$E'\{\sum_{k=0}^{K-1} [\mathcal{A}'\eta(x_k')|x_0\} = E'\{\eta(x_K)]|x_0\} - \eta(x_0). \tag{20}$$

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 (20), we obtain the performance difference formula:

$$\eta'(x_0) - \eta(x_0) = \lim_{K\to\infty} \sum_{k=0}^{K-1} E'\{(\mathcal{A}'\eta + f')(x_k')|x_0\}. \tag{21}$$

B. Optimal Policy

Based on the performance difference formula (21), we have the following optimality condition.

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

$$\mathcal{A}^{\mathbf{u}}\eta^{\mathbf{u}^*} + f^{\mathbf{u}} \geq 0 = \mathcal{A}^{\mathbf{u}^*}\eta^{\mathbf{u}^*} + f^{\mathbf{u}^*}, \forall \mathbf{u} \in \mathcal{C}. \tag{22}$$

From (22), the optimality equation is:

$$\min_{\mathbf{u}\in\mathcal{C}}\{\mathcal{A}^{\mathbf{u}}\eta^{\mathbf{u}^*} + f^{\mathbf{u}}\} = 0. \tag{23}$$

Proof: Firstly, the "if" part follows directly from the performance difference formula (21), and the Poisson equation (18).

Next, we prove the "only if" part: Let $\mathbf{u}^*$ be an optimal policy. We need to prove that (22) holds. Suppose that this is not true. Then, there must exist one policy, denoted as $\mathbf{u}'$, such that (22) 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 (23). 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), (9), and (11), we have

$$\begin{aligned}(P^{\mathbf{u}}\eta^{\mathbf{u}_k})(x) =& (a_1\hat{\mathbf{G}} + a_2\bar{\mathbf{G}})(A + \mathbf{B}\hat{\mathbf{K}})^2x^2\mathbf{1}_{x\geq 0} \\ &+ (a_1\hat{\mathbf{G}} + a_2\bar{\mathbf{G}})(A - \mathbf{B}\bar{\mathbf{K}})^2x^2\mathbf{1}_{x<0}.\end{aligned} \tag{24}$$

where a_1 and a_2 satisfy equations (12) and (13), respectively.

Then, from (4) and (24), 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 $\hat{C} = A + \mathbf{B}\mathbf{K}$, and $\bar{C} = A - \mathbf{B}\mathbf{K}$.

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 (23) in Theorem 1, and therefore is an optimal control.

Therefore, we can obtain the optimal policy as follows,

$$\mathbf{u}^*(x) = \hat{\mathbf{K}}^*x\mathbf{1}_{x\geq 0} - \bar{\mathbf{K}}^*x\mathbf{1}_{x<0}, \tag{25}$$

where

$$\hat{\mathbf{K}}^* = \arg\min_{\mathbf{K}\in\mathcal{K}}[a_1\hat{C}^2\hat{\mathbf{G}}^* + a_2\hat{C}^2\bar{\mathbf{G}}^* + Q + \mathbf{K}^T\mathbf{R}\mathbf{K}], \tag{26}$$

$$\bar{\mathbf{K}}^* = \arg\min_{\mathbf{K}\in\mathcal{K}}[a_1\bar{C}^2\hat{\mathbf{G}}^* + a_2\bar{C}^2\bar{\mathbf{G}}^* + Q + \mathbf{K}^T\mathbf{R}\mathbf{K}]. \tag{27}$$

Moreover,

$$\hat{\mathbf{G}}^* = \min_{\mathbf{K}\in\mathcal{K}}\{a_1\hat{C}^2\hat{\mathbf{G}}^* + a_2\hat{C}^2\bar{\mathbf{G}}^* + Q + \mathbf{K}^T\mathbf{R}\mathbf{K}\}, \tag{28}$$

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

The original problem (3) is transferred to two auxiliary optimization problems (26) and (27). Under the optimal control $\mathbf{u}^*$ in (25), the closed-loop system (8) is L^2-asymptotically stable. From (16), with the initial condition $x_0 = x$, we know the optimal total reward performance of this system is

$$\eta^*(x) = \hat{\mathbf{G}}^*x^2\mathbf{1}_{x\geq 0} + \bar{\mathbf{G}}^*x^2\mathbf{1}_{x<0}, \tag{30}$$

where $\hat{\mathbf{G}}^*$ and $\bar{\mathbf{G}}^*$ satisfy (28) and (29), 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. Besides, [6] also provides some algorithms for calculating the optimal policy.

IV. Simulation Example

In this section, we use a simple example to illustrate the optimal policy for the constrained LQ control problem (3).

We consider a stochastic LQ system with $x_0 = 15$, $m = 3$, $A = 0.8$, and $B = (-0.35, 0.18, 0.25)$. The cost matrices are

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

For time $l = 0, 1, \cdots$, the variance of 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 $\hat{\mathbf{K}}^* = (0.574, 0, 0)^T$, $\bar{\mathbf{K}}^* = (0, 0.250, 0.270)^T$, $\hat{\mathbf{G}}^* = 2.773$ and $\bar{\mathbf{G}}^* = 3.473$. Furthermore, the optimal total reward performance is $\eta^*(x_0) = \hat{\mathbf{G}}^*x_0^2\mathbf{1}_{x_0\geq 0} + \bar{\mathbf{G}}^*x_0^2\mathbf{1}_{x_0<0} = 623.987$. It can be observed that x_l^* converges to 0 fast and this closed loop system is asymptotically stable.

V. Conclusions

In this paper, we apply the direct-comparison based optimization approach to study the total rewards optimization of the discrete-time stochastic linear-quadratic control problem with conic control constraints in infinite horizon. We derive the performance difference formula by utilizing the state separation property. Based on it, 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. The results can be extended to the cases of non-Gaussian noises and average rewards easily. 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. Next step, we can investigate continuous cases. Besides, since 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, financial derivative pricing, in our further research.

References

[1] Y. Hu and X. Y. Zhou, "Constrained stochastic LQ control with random coefficients, and application to portfolio selection," *SIAM Journal on Control and Optimization*, vol. 44, no. 2, pp. 444–446, 2005.

[2] R. E. Kalman, "Contributions to the theory of optimal control," *Bol. Soc. Mat. Mexicana*, vol. 5, no. 2, pp. 102–119, 1960.

[3] B. D. Anderson and J. B. Moore, *Optimal Control: Linear Quadratic Methods*, Courier Corporation, 2007.

[4] J. Yong, "Linear-quadratic optimal control problems for mean-field stochastic differential equations," *SIAM Journal on Control and Optimization*, vol. 51, no. 4, pp. 2809–2838, 2013.

[5] J. J. Gao and D. Li, "Cardinality constrained linear quadratic optimal control," *IEEE Trans. Autom. Control*, vol. 56, no. 8, pp. 1936–1941, 2011.

[6] W. P. Wu, J. J. Gao, D. Li, and Y. Shi, "Explicit solution for constrained stochastic linear-quadratic control with multiplicative noise," *IEEE Trans. Autom. Control*, submitted for publication.

[7] X. R. Cao, *Stochastic Learning and Optimization: A Sensitivity-Based Approach*. New York: Springer, 2007.

[8] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, New York: Wiley, 1994.

[9] K. J. Zhang, Y. K. Xu, X. Chen, and X. R. Cao, "Policy iteration based feedback control," *Automatica*, vol. 44, no. 2, pp. 1055–1061, 2008.

[10] X. R. Cao, "Stochastic feedback control with one-dimensional degenerate diffusions and nonsmooth value functions," *IEEE Trans. Autom. Control*, vol. 62, no. 12, pp. 6136–6151, 2017.

[11] X. R. Cao, and X. W. Wan, "Sensitivity analysis of nonlinear behavior with distorted probability," *Mathematical Finance*, vol. 27, no. 1, pp. 115–150, 2017.

[12] L. Xia, "Mean-variance optimization of discrete time discounted Markov decision processes," *Automatica*, vol. 88, pp. 76–82, 2018.

[13] X. S. Ye, R. B. Xue, J. J. Gao, and X. R. Cao, "Optimization in curbing risk contagion among financial institutes," *Automatica*, submitted for publication.