Policy Optimization for Markovian Jump Linear Quadratic Control: Gradient Method and Global Convergence

Joao Paulo Jansch-Porto, Bin Hu, and Geir E. Dullerud

***Abstract*—Recently, policy optimization has received renewed attention from the control community due to various applications in reinforcement learning tasks. In this paper, we investigate the global convergence of the gradient method for quadratic optimal control of discrete-time Markovian jump linear systems (MJLS). First, we study the optimization landscape of direct policy optimization for MJLS, with static state feedback controllers and quadratic performance costs. Despite the non-convexity of the resultant problem, we are still able to identify several useful properties such as coercivity, gradient dominance, and smoothness. Based on these properties, we prove that the gradient method converges to the optimal state feedback controller for MJLS at a linear rate if initialized at a controller which is mean-square stabilizing. This work brings new insights for understanding the performance of the policy gradient method on the Markovian jump linear quadratic control problem.**

***Index Terms*—Markovian jump linear systems, optimal control, policy gradient methods, reinforcement learning.**

I. Introduction

Recently, reinforcement learning (RL) [1] has achieved impressive performance on continuous control tasks such as locomotion [2] and robotic hand manipulation [3]. One main algorithmic framework for such RL applications is policy optimization [4]. Specifically, policy-based RL methods including the policy gradient method [5], natural policy gradient [6], TRPO [7], natural AC [8], and PPO [9], have been widely used in various control tasks. These methods enable flexible policy parameterizations and optimize control performance directly.

Although policy-based RL methods have shown great promise in addressing complex control tasks, the selection and tuning of these methods have not been fully understood [10], [11]. This has motivated a recent research trend focusing on understanding the performances of policy optimization algorithms on simplified benchmarks such as linear quadratic regulator (LQR) [12]–[22], linear robust control [23]–[25], and linear control of Lur'e systems [26]. Notice that even for LQR, directly optimizing over the policy space leads to a non-convex constrained problem. Nevertheless, one can still prove the global convergence of policy gradient methods on the LQR problem by exploiting properties such as gradient dominance, almost smoothness, and coercivity [12], [13]. This provides a good sanity check for applying policy optimization to more advanced control applications.

Built upon the good progress on understanding policy-based RL for linear time-invariant (LTI) systems, this paper moves one step further and presents new theoretical results on policy optimization of Markov jump linear systems (MJLS) [27]. MJLS form an important class of hybrid dynamical systems that find many applications in control [28]–[33] and machine learning [34], [35]. The research on MJLS has great practical value while in the mean time also provides new interesting theoretical questions. Different from the LTI case, the state/input matrices of a Markov jump linear system are functions of a jump parameter sampled from an underlying Markov chain. Controlling unknown MJLS poses many new challenges over traditional LQR due to the appearance of this Markov jump parameter, and it is the coupling effect between the state/input matrices and the jump parameter distribution that causes the main difficulty. To this end, the optimal control of MJLS provides a meaningful benchmark for further understanding of policy-based RL algorithms.

However, the theoretical properties of policy-based RL methods on discrete-time MJLS have been overlooked in the existing literature [36]–[39]. In this paper, we make one step towards bridging this gap. Specifically, we develop new convergence theory for direct policy optimization of MJLS. Despite the non-convexity of the resultant policy search problem, we are still able to identify several useful properties such as coercivity, gradient dominance, and smoothness. Then we use these identified properties to prove that the gradient method converges to the optimal state feedback controller for MJLS at a linear rate if a stabilizing initial controller is used.

Our paper generalizes the convergence theory for LTI policy optimization [12], [13], [20] to the MJLS case. This extension is nontrivial, and heavily relies on the operator-theoretic stability arguments used in the MJLS literature [27]. Our paper also expands on the previous results published by the authors in a conference paper [40], and has made significant extensions in identifying the smoothness property and analyzing the gradient descent method in the MJLS setting. Our work serves as an important step toward understanding the theoretical aspects of policy-based RL methods for MJLS control. There is a follow-up work [41] which extended our convergence theory of the gradient method to the model-free policy gradient setting. When the models are unknown, the gradient method can still be implemented using zeroth-order optimization techniques and yield global convergence guarantees [41]. The sample complexity analysis in [41] heavily relies on the cost properties identified in our paper.

II. Background and Problem Formulation

A. Notation

We denote the set of real numbers by $\mathbb{R}$. Let A be a matrix, then we use the notation A^T, $\|A\|$, $\operatorname{tr}(A)$, $\sigma_{\min}(A)$, and $\rho(A)$ to denote its transpose, maximal singular value, trace, minimum singular value, and spectral radius, respectively. Given matrices $\{D_i\}_{i=1}^m$, let $\operatorname{diag}(D_1, \dots, D_m)$ denote the block diagonal matrix whose (i,i)-th block is D_i. The Kronecker product of matrices A and B is denoted as $A \otimes B$. We use $\operatorname{vec}(A)$ to denote the vectorization of matrix A. We indicate when a symmetric matrix Z is positive definite or positive semidefinite matrices by $Z \succ 0$ and $Z \succeq 0$, respectively. Given a function f, we use df to denote its total derivative [42].

We now introduce some specific matrix spaces and notation motivated from the MJLS literature [27]. Let $\mathbb{M}^N_{n\times m}$ denote the space made up of all N-tuples of real matrices $V = (V_1, \dots, V_N)$ with $V_i \in \mathbb{R}^{n\times m}, i \in \mathbb{N}$. For simplicity, we write $\mathbb{M}^N$ in place of $\mathbb{M}^N_{n\times m}$ when the dimensions n and m are clear from context. For

J. P. Jansch-Porto is with the Department of Mechanical Science and Engineering, University of Illinois at Urbana-Champaign, Email: janschp2@illinois.edu.

B. Hu is with the Coordinated Science Laboratory (CSL) and the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Email: binhu7@illinois.edu.

G. E. Dullerud is with the Coordinated Science Laboratory (CSL) and the Department of Mechanical Science and Engineering, University of Illinois at Urbana-Champaign, Email: dullerud@illinois.edu.

J. P. Jansch-Porto and G. Dullerud are funded by NSF under the grant ECCS 19-32735. B. Hu is funded by the NSF award CAREER-2048168.

$V = (V_1, \dots, V_N) \in \mathbb{M}^N$, we define

$$\|V\|_1 := \sum_{i\in\Omega} \|V_i\|, \quad \|V\|_2^2 := \sum_{i=1}^{N} \operatorname{tr}\left(V_i^T V_i\right),$$
$$\|V\|_{\max} := \max_{i=1,\dots,N} \|V_i\|, \quad \Lambda_{\min}(V) := \min_{i=1,\dots,N} \sigma_{min}(V_i).$$

Clearly, we have $\|V\|_{\max} \leq \|V\|_1 \leq \|V\|_2$. For $V, S \in \mathbb{M}^N$, their inner product is defined as

$$\langle V,\, S\rangle := \sum_{i=1}^{N} \operatorname{tr}\left(V_i^T S_i\right)$$

Notice both V and S are sequences of matrices. It is also convenient to define $V + S := (V_1 + S_1, \dots, V_N + S_N)$, $VS := (V_1 S_1, \dots, V_N S_N)$, $V^T := (V_1^T, \dots, V_N^T)$, and $V^{-1} := (V_1^{-1}, \dots, V_N^{-1})$. We say that $V \succ S$ if $V_i - S_i \succ 0$ for $i = 1, \dots, N$.

B. Markovian Jump Linear Quadratic Control

In this paper, we consider the optimal control of the following discrete-time Markovian jump linear system (MJLS):

$$x_{t+1} = A_{\omega(t)} x_t + B_{\omega(t)} u_t \tag{1}$$

where $x_t \in \mathbb{R}^d$ is the system state, and $u_t \in \mathbb{R}^k$ corresponds to the control action. The system matrices $A_{\omega(t)} \in \mathbb{R}^{d\times d}$ and $B_{\omega(t)} \in \mathbb{R}^{d\times k}$ depend on the switching parameter $\omega(t)$, which takes values on $\Omega := \{1, \dots, N_s\}$. We will denote $A = (A_1, \dots, A_{N_s}) \in \mathbb{M}^{N_s}_{d\times d}$ and $B = (B_1, \dots, B_{N_s}) \in \mathbb{M}^{N_s}_{d\times k}$.

The jump parameter $\{\omega(t)\}_{t=0}^{\infty}$ is assumed to form a time-homogeneous Markov chain whose transition probability is given as

$$p_{ij} = \mathbb{P}\left(\omega(t+1) = j | \omega(t) = i\right). \tag{2}$$

Let $\mathcal{P}$ denote the probability transition matrix whose (i,j)-th entry is p_{ij}. The initial distribution of $\omega(0)$ is given by $\pi = \begin{bmatrix}\pi_1 & \cdots & \pi_{N_s}\end{bmatrix}^T$. Obviously, we have $p_{ij} \geq 0$, $\sum_{j=1}^{N_s} p_{ij} = 1$, and $\sum_{i\in\Omega} \pi_i = 1$. We further assume that system (1) is mean-square stabilizable[1].

Our control design objective is to choose the actions $\{u_t\}_{t=0}^{\infty}$ to minimize the following quadratic cost function

$$C = \mathbb{E}_{x_0\sim\mathcal{D},\omega_0\sim\pi}\left[\sum_{t=0}^{\infty} x_t^T Q_{\omega(t)} x_t + u_t^T R_{\omega(t)} u_t\right], \tag{3}$$

where $\mathcal{D}$ denotes the initial state distribution. For simplicity, it is assumed that $Q = (Q_1, \dots, Q_{N_s}) \succ 0$, $R = (R_1, \dots, R_{N_s}) \succ 0$, $\pi_i > 0$, and $\mathbb{E}_{x_0\sim\mathcal{D}}\left[x_0 x_0^T\right] \succ 0$. The assumptions on π and $\mathbb{E}_{x_0\sim\mathcal{D}}\left[x_0 x_0^T\right]$ indicate that there is a chance of starting from any mode i and the covariance of the initial state is full rank. These assumptions can be somehow informally thought as the persistently excitation condition in the system identification literature and are quite standard for learning-based control. The above problem can be viewed as the MJLS counterpart of the standard LQR problem, and hence is termed as the "MJLS LQR problem." It is known that the optimal cost for the MJLS LQR problem can be achieved by a linear state feedback of the form

$$u_t = -K_{\omega(t)} x_t \tag{4}$$

with $K = (K_1, \dots, K_{N_s}) \in \mathbb{M}^{N_s}_{k\times d}$. Combining the linear policy (4) with (1), we obtain the closed-loop dynamics:

$$x_{t+1} = \left(A_{\omega(t)} - B_{\omega(t)} K_{\omega(t)}\right) x_t = \Gamma_{\omega(t)} x_t. \tag{5}$$

with $\Gamma = (\Gamma_1, \dots, \Gamma_{N_s}) \in \mathbb{M}^{N_s}_{d\times d}$. Note that using this formulation, we can write the cost (3) as

$$C = \mathbb{E}_{x_0\sim\mathcal{D},\omega_0\sim\pi}\left[\sum_{t=0}^{\infty} x_t^T \left(Q_{\omega(t)} + K_{\omega(t)}^T R_{\omega(t)} K_{\omega(t)}\right) x_t\right].$$

The optimal controller to the above MJLS LQR problem can be computed by solving a system of coupled Algebraic Riccati Equations (AREs) [43]. Specifically, define the operator $\mathcal{E} : \mathbb{M}^{N_s}_{d\times d} \to \mathbb{M}^{N_s}_{d\times d}$ as $\mathcal{E}(V) := (\mathcal{E}_1(V), \dots, \mathcal{E}_{N_s}(V))$ where $V = (V_1, \dots, V_{N_s}) \in \mathbb{M}^{N_s}_{d\times d}$ and $\mathcal{E}_i(V) := \sum_{j=1}^{N_s} p_{ij} V_j$. Let $P = (P_1, \dots, P_{N_s})$ be the unique positive definite solution to the following AREs:

$$P = Q + A^T\mathcal{E}(P)A - A^T\mathcal{E}(P)B\times \left(R + B^T\mathcal{E}(P)B\right)^{-1} B^T\mathcal{E}(P)A. \tag{6}$$

Then, it is known that the optimal controller is given by

$$K^* = \left(R + B^T\mathcal{E}(P)B\right)^{-1} B^T\mathcal{E}(P)A. \tag{7}$$

Notice that the existence of such a controller is guaranteed by the stabilizability assumption. In this paper, we will revisit the above MJLS LQR problem from a policy optimization perspective.

C. Policy Optimization for LTI Systems

Before proceeding to policy optimization of MJLS, we briefly review some relevant results for LTI systems [12]. Consider the LTI system $x_{t+1} = Ax_t + Bu_t$, where $A \in \mathbb{R}^{d\times d}$, and $B \in \mathbb{R}^{d\times k}$. Let u_t be determined by a static state feedback controller, i.e. $u_t = -Kx_t$. We adopt the following standard quadratic cost function

$$\begin{aligned} C(K) &= \mathbb{E}_{x_0\sim\mathcal{D}}\left[\sum_{t=0}^{\infty} x_t^T Q x_t + u_t^T R u_t\right] \\ &= \mathbb{E}_{x_0\sim\mathcal{D}}\left[\sum_{t=0}^{\infty} x_t^T (Q + K^T R K) x_t\right], \end{aligned} \tag{8}$$

which is equal to $\mathbb{E}_{x_0\sim\mathcal{D}}\left[x_0^T P^K x_0\right]$ with P^K being the solution to the Lyapunov equation $P^K = Q + K^T R K + (A - BK)^T P^K (A - BK)$. The following gradient formula [12], [44] is also well known

$$\nabla C(K) = 2\left(\left(R + B^T P^K B\right) K - B^T P^K A\right)\Sigma_K,$$

where $\Sigma_K = \mathbb{E}_{x_0\sim\mathcal{D}}\left[\sum_{t=0}^{\infty} x_t x_t^T\right]$. Based on this gradient formula, there exists a unique K^* such that $\nabla C(K^*) = 0$ if $\mathbb{E}_{x_0\sim\mathcal{D}}\left[x_0 x_0^T\right]$ is full rank [12]. In addition, one can optimize (8) using the gradient method $K' \leftarrow K - \eta\nabla C(K)$, or the natural policy gradient method $K' \leftarrow K - \eta\nabla C(K)\Sigma_K^{-1}$. In [12], it has been shown that these methods are guaranteed to converge to K^* linearly if a stabilizing initial policy is used. One advantage of these gradient-based methods is that they can be implemented in a model-free manner. More discussions on these methods and their model-free variants can be found in [12].

D. Problem Setup: Policy Optimization for MJLS

In this section, we reformulate the MJLS LQR problem as a policy optimization problem. Since the optimal cost for the MJLS LQR problem can be achieved by a linear state feedback controller, it is reasonable to confine the policy search to the class of linear state feedback policies. Hence we set $K = (K_1, \dots, K_{N_s})$, and the control action is determined as $u_t = -K_{\omega(t)} x_t$. This leads to the following policy optimization problem whose decision variable is K.

[1] The mean square stability of MJLS is reviewed in sequel.

Problem: Policy Optimization for MJLS.

minimize: *cost $C(K)$, given in (3)*
subject to: *state dynamics, given in (1)*
control actions, given in (4)
transition probabilities, given in (2)
stability constraint, K stabilizing (1) in the mean square sense.

When $N_s = 1$, the above problem reduces to the policy optimization for LTI systems [12]. We want to emphasize that the above problem is indeed a constrained optimization problem. Recall that given K, the resultant closed-loop MJLS (5) is mean square stable (MSS) if for any initial condition $x_0 \in \mathbb{R}^d$ and $\omega(0) \in \Omega$, one has $\mathbb{E}\left[x_t x_t^T\right] \to 0$ as $t \to \infty$ [27]. Since it is assumed $\mathbb{E}_{x_0\sim\mathcal{D}}\left[x_0 x_0^T\right] \succ 0$, we can trivially apply the well-known equivalence between mean square stability and stochastic stability for MJLS [27] to show that $C(K)$ is finite if and only if K stabilizes the closed-loop dynamics in the mean square sense. Therefore, the feasible set of the above policy optimization problem consists of all K stabilizing the closed-loop dynamics (5) in the mean square sense. For simplicity, we denote this feasible set as $\mathbb{K}$. For $K \in \mathbb{K}$, $C(K)$ can be calculated as

$$\begin{aligned} C(K) &= \mathbb{E}_{x_0\sim\mathcal{D},\omega_0\sim\pi}\left[x_0^T P^K_{\omega(0)} x_0\right] \\ &= \mathbb{E}_{x_0\sim\mathcal{D}}\left[x_0^T \left(\sum_{i\in\Omega} \pi_i P_i^K\right) x_0\right], \end{aligned} \tag{9}$$

where $P^K = (P_1^K, \ldots, P_{N_s}^K) \in \mathbb{M}^{N_s}_{d\times d}$ and each P_i^K is solved via the following coupled Lyapunov equations:

$$P_i^K = Q_i + K_i^T R_i K_i + (A_i - B_i K_i)^T \mathcal{E}_i(P^K)(A_i - B_i K_i). \tag{10}$$

The goal for policy optimization is to apply iterative gradient-based methods to search for the cost-minimizing element K^* within the feasible set $\mathbb{K}$. A fundamental question is how to check whether $K \in \mathbb{K}$ for any given K. There are several ways to do this, and we give a brief review here. We need to introduce a few operators which are standard in the MJLS literature. Specifically, for any $V \in \mathbb{M}^{N_s}_{d\times d}$, we define $\mathcal{T}(V) = (\mathcal{T}_1(V), \ldots, \mathcal{T}_{N_s}(V)) \in \mathbb{M}^{N_s}_{d\times d}$, where $\mathcal{T}_j(V)$ is computed as

$$\mathcal{T}_j(V) := \sum_{i\in\Omega} p_{ij}(A_i - B_i K_i)V_i(A_i - B_i K_i)^T.$$

Recall that $\mathcal{E}_i(V) := \sum_{j=1}^{N_s} p_{ij} V_j$. We can also define $\mathcal{L}(V) = (\mathcal{L}_1(V), \ldots, \mathcal{L}_{N_s}(V)) \in \mathbb{M}^{N_s}_{d\times d}$, where $\mathcal{L}_i(V)$ is given as

$$\mathcal{L}_i(V) := (A_i - B_i K_i)^T \mathcal{E}_i(V)(A_i - B_i K_i).$$

The following property of $\mathcal{E}_i$ is quite useful

$$\|\mathcal{E}_i(V)\| \le \sum_{j\in\Omega} p_{ij}\|V_i\| \le \|V\|_{\max}\left(\sum_{j\in\Omega} p_{ij}\right) = \|V\|_{\max}. \tag{11}$$

It is also easy to check that both $\mathcal{T}$ and $\mathcal{L}$ are Hermitian and positive operators. From [27], we also know $\mathcal{T}$ is the adjoint operator of $\mathcal{L}$. The operator $\mathcal{T}$ is useful in describing the covariance propagation of the MJLS (5). Specifically, if we define $X(t) = (X_1(t), \ldots, X_{N_s}(t))$ with $X_i(t) := \mathbb{E}\left[x_t x_t^T \mathbf{1}_{\omega(t)=i}\right]$, then we have $X(t+1) = \mathcal{T}(X(t))$. In addition, we know $\sum_{t=0}^\infty X(t)$ exists if $K \in \mathbb{K}$. We denote this limit as $\mathbf{X}^K$ and we have

$$\mathbf{X}^K = \sum_{t=0}^\infty \mathcal{T}^t(X(0)). \tag{12}$$

The operator $\mathcal{L}$ is useful for value computation, since we have $P^K = \mathcal{L}(P^K) + Q + K^T R K$ (or equivalently $P^K = \sum_{t=0}^\infty \mathcal{L}^t(Q + K^T R K)$) for any $K \in \mathbb{K}$. Also notice $\mathcal{L}$ is actually a linear operator and has a matrix representation $\mathcal{A} := \operatorname{diag}\left(\Gamma_i^T \otimes \Gamma_i^T\right)(\mathcal{P} \otimes I_{N_s^2})$ where $\Gamma_i = A_i - B_i K_i$ (see Proposition 3.4 in [27] for more details). Now we are ready to present the following well-known result which can be used to check whether K is in $\mathbb{K}$ or not.

Proposition 1 ([27]). *The following assertions are equivalent:*

1) *System (5) is MSS.*
2) *$\rho(\mathcal{A}) < 1$.*
3) *For any $S \in \mathbb{M}^{N_s}_{n\times n}$, $S \succ 0$, there exists a unique $V \in \mathbb{M}^{N_s}_{n\times n}$, $V \succ 0$, such that $V - \mathcal{T}(V) = S$.*
4) *There exists $V \succ 0 \in \mathbb{M}^{N_s}_{n\times n}$ such that $V - \mathcal{T}(V) \succ 0$.*

The results above also hold when replacing $\mathcal{T}$ by $\mathcal{L}$.

Based on the above result, a few basic properties of $\mathbb{K}$ can be obtained. Clearly, we have $\mathbb{K} := \{K \in \mathbb{M}^{N_s}_{k\times d} : \rho(\mathcal{A}) < 1\}$. Since $\rho(\mathcal{A})$ is a continuous function of K, we know $\mathbb{K}$ is an open set and $\mathbb{K}^c$ is a closed set. The boundary of the set $\mathbb{K}$ can also be formally specified as $\partial\mathbb{K} := \{K \in \mathbb{M}^{N_s}_{k\times d} : \rho(\mathcal{A}) = 1\}$.

Finally, it is worth mentioning that both $\mathcal{L}$ and $\mathcal{T}$ depend on K. Occasionally, we will use the notation $\mathcal{L}^K$ and $\mathcal{T}^K$ when there is a need to emphasize the dependence of these operators on K.

III. Optimization Landscape and Cost Properties

In this section, we study the optimization landscape of the MJLS LQR problem and identify several useful properties of $C(K)$. Based on Lemma 1 in [40], the cost (9) is continuously differentiable with respect to K, and the gradient $\nabla C(K)$ can be calculated as

$$\nabla C(K) = 2L^K \mathbf{X}^K \tag{13}$$

where $L^K = (L_1^K, \ldots, L_{N_s}^K) \in \mathbb{M}^{N_s}_{k\times d}$ and L_i^K is given by

$$L_i^K = \left(R_i + B_i^T \mathcal{E}_i(P^K) B_i\right) K_i - B_i^T \mathcal{E}_i(P^K) A_i. \tag{14}$$

Moreover, $\mathbf{X}^K$ in the above gradient formula is given by (12). Since K is a tuple of real matrices, we have $\nabla C(K) \in \mathbb{M}^{N_s}_{k\times d}$. Next, we present an explicit formula for the Hessian of the cost. To avoid tensors, we restrict analysis with the quadratic form of the Hessian $\nabla^2 C(K)[E,E]$ on a matrix sequence $E \in \mathbb{M}^{N_s}_{k\times d}$.

Lemma 1. *For $K \in \mathbb{K}$, the Hessian of the MJLS LQR cost $C(K)$ applied to a direction $E \in \mathbb{M}^{N_s}_{k\times d}$ is given by*

$$\begin{aligned} \nabla^2 C(K)[E,E] = 2\langle (R + B^T \mathcal{E}(P^K) B) E \mathbf{X}^K,\, E\rangle \\ - 4\langle B^T \mathcal{E}((P^K)'[E]) \varGamma \mathbf{X}^K,\, E\rangle, \end{aligned} \tag{15}$$

where

$$(P^K)'[E] = \sum_{t=0}^\infty \mathcal{L}^t(E^T L^K + (L^K)^T E). \tag{16}$$

Proof. Recall that $\varGamma_i := A_i - B_i K_i$ with $\varGamma = (\varGamma_1, \ldots, \varGamma_{N_s})$. Applying the Taylor series expansion about E [45], we have that the quadratic form of the hessian $\nabla^2 C(K)$ on E is given by

$$\nabla^2 C(K)[E,E] = \left.\frac{d^2}{dt^2}\right|_{t=0} C(K + tE). \tag{17}$$

Writing (9) as $C(K) = \langle P^K, X(0)\rangle$, we then have

$$\nabla^2 C(K)[E,E] = \langle \frac{d^2}{dt^2}\Big|_{t=0} P^{K+tE}, X(0)\rangle. \tag{18}$$

Denote $(P^K)'[E] := \frac{d}{dt}\big|_{t=0} P^{K+tE}$. The following equation holds

$$\begin{aligned}(P^K)'[E] = \sum_{t=0}^{\infty} \mathcal{L}^t\Big(&E^T(RK - B^T\mathcal{E}(P^K)\mathit{\Gamma}) \\ &+ (RK - B^T\mathcal{E}(P^K)\mathit{\Gamma})^T E\Big).\end{aligned}$$

We can show that

$$\frac{d^2}{dt^2}\Big|_{t=0} P^{K+tE} = \sum_{t=0}^{\infty} \mathcal{L}^t(S), \tag{19}$$

where S is given as

$$\begin{aligned}S = -2\left(E^T B^T \mathcal{E}((P^K)'[E])\mathit{\Gamma} + \mathit{\Gamma}\mathcal{E}((P^K)'[E])BE\right) \\ + 2\, E^T(R + B^T\mathcal{E}(P^K)B)E.\end{aligned} \tag{20}$$

Since $\mathcal{T}$ is the adjoint operator of $\mathcal{L}$, we have

$$\begin{aligned}\nabla^2 C(K)[E,E] &= \langle \sum_{t=0}^{\infty} \mathcal{L}^t(S), X(0)\rangle \\ &= \langle S, \sum_{t=0}^{\infty} \mathcal{L}^t(X(0))\rangle = \langle S, \mathbf{X}^K\rangle.\end{aligned}$$

Plugging (20) into the above, we get

$$\begin{aligned}\nabla^2 C(K)[E,E] = 2\langle E^T(R + B^T\mathcal{E}(P^K)B)E, \mathbf{X}^K\rangle \\ - 2\langle E^T B^T \mathcal{E}((P^K)'[E])\mathit{\Gamma}, \mathbf{X}^K\rangle \\ - 2\langle \mathit{\Gamma}\mathcal{E}((P^K)'[E])BE, \mathbf{X}^K\rangle \\ = 2\langle E^T(R + B^T\mathcal{E}(P^K)B)E, \mathbf{X}^K\rangle \\ - 4\langle E^T B^T \mathcal{E}((P^K)'[E])\mathit{\Gamma}, \mathbf{X}^K\rangle.\end{aligned}$$

We can get the desired result by noting that each block in $\mathbf{X}^K$ is symmetric and using the cyclic property of the trace. □

Optimization Landscape for MJLS. Now we discuss the optimization landscape for the MJLS LQR problem. Notice that LTI systems are just a special case of MJLS. Since policy optimization for quadratic control of LTI systems is non-convex, the same is true for the MJLS case. By examining the gradient formula (13), it becomes clear that as long as $\mathbb{E}_{x_0\sim\mathcal{D}}\left[x_0 x_0^T\right]$ is full rank and $\pi_i > 0$ for all i, any stationary point given by $\nabla C(K) = 0$ has to satisfy

$$L_i^K = \left(R_i + B_i^T\mathcal{E}_i(P^K)B_i\right)K_i - B_i^T\mathcal{E}_i(P^K)A_i = 0.$$

Substituting the above equation into (10) leads to the global solution K^* defined by the coupled AREs (6), and hence the only stationary point is the global optimal solution. When the initial mode is sufficiently random, i.e. $\pi_i > 0$ for all i, the optimization landscape for the MJLS case becomes quite similar to the classic LQR case. Based on such similarity, it is reasonable to expect that gradient-based methods will work well in the MJLS LQR setting despite the non-convex nature of the problem. Compared with the LTI case, the characterization of $\mathbb{K}$ is more complicated for MJLS. Hence one main technical issue is how to show gradient-based methods can handle the feasibility constraint $K \in \mathbb{K}$ without using projection.

Key Properties of the MJLS LQR Cost. To analyze the performance of gradient-based methods for the MJLS LQR problem, a few key properties of $C(K)$ will be used. By assumption, we have $\mu := \min_{i\in\Omega}(\pi_i)\,\sigma_{\min}\left(\mathbb{E}_{x_0\sim\mathcal{D}}\left[x_0x_0^T\right]\right) > 0$. Then, we can identify several key properties of $C(K)$ as follows.

Lemma 2. *The cost (9) satisfies the following properties:*

1) *Coercivity: The cost function C is coercive in the sense that for any sequence $\{K^l\}_{l=1}^{\infty} \subset \mathbb{K}$ we have*

$$C(K^l) \to +\infty$$

if either $\|K^l\|_2 \to +\infty$, or K^l converges to an element K in the boundary $\partial\mathbb{K}$.

2) *Almost smoothness: Given elements $K, K' \in \mathbb{K}$, the cost function $C(K)$ defined in (3) satisfies*

$$\begin{aligned}&C(K') - C(K) \\ &= \sum_{i\in\Omega}\Big(-2\,\mathrm{tr}\Big(\mathbf{X}_i^{K'}\Delta K_i^T L_i^K\Big) \\ &\quad + \mathrm{tr}\Big(\mathbf{X}_i^{K'}\Delta K_i^T\left(R_i + B_i^T\mathcal{E}_i(P^K)B_i\right)\Delta K_i\Big)\Big) \\ &= -2\langle \Delta K^T L^K, \mathbf{X}^{K'}\rangle + \langle \Delta K^T\Psi\Delta K, \mathbf{X}^{K'}\rangle,\end{aligned}$$

where $\Delta K_i = (K_i - K_i')$ and $\Psi := R + B^T\mathcal{E}(P^K)B$.

3) *Gradient dominance: Given the optimal policy K^*, the following sequence of inequalities holds for any $K \in \mathbb{K}$:*

$$\begin{aligned}C(K) - C(K^*) &\le \|\mathbf{X}^{K^*}\|_{\max}\langle \Psi^{-1}L^K, L^K\rangle \\ &\le \frac{\|\mathbf{X}^{K^*}\|_{\max}}{\Lambda_{\min}(R)}\|L^K\|_2^2 \\ &\le \frac{\|\mathbf{X}^{K^*}\|_{\max}}{4\mu^2\Lambda_{\min}(R)}\|\nabla C(K)\|_2^2.\end{aligned}$$

4) *Compactness of the sublevel sets: The sublevel set defined as $\mathbb{K}_\alpha := \{K \in \mathbb{K} : C(K) \le \alpha\}$ is compact for every $\alpha \ge C(K^*)$.*

5) *Smoothness on the sublevel sets: For any sublevel set $\mathbb{K}_\alpha$, choose the smoothness constant as*

$$L = 2\left(\|R\|_{\max} + \|B\|_{\max}^2\left(1 + \frac{2\xi}{\|B\|_{\max}}\right)\frac{\alpha}{\mu}\right)\frac{\alpha}{\Lambda_{\min}(Q)},$$

where ξ is calculated as

$$\xi = \frac{1}{\Lambda_{\min}(Q)}\left(\frac{1 + \|B\|_{\max}^2}{\mu}\alpha + \|R\|_{\max}\right) - 1.$$

Then for any $K \in \mathbb{K}_\alpha$, we have $\|\nabla^2 C(K)\| \le L$. In addition, for any (K, K') satisfying $tK + (1-t)K' \in \mathbb{K}_\alpha$ $\forall t \in [0,1]$, the following inequality holds

$$C(K') \le C(K) + \langle \nabla C(K), K' - K\rangle + \frac{L}{2}\|K' - K\|_2^2. \tag{21}$$

Proof. To prove Statement 1, first notice that we have

$$\begin{aligned}C(K^l) &\ge \mathbb{E}_{x_0\sim\mathcal{D}}\left[\sum_{i\in\Omega}\pi_i x_0^T(Q_i + (K_i^l)^T R_i K_i^l)x_0\right] \\ &\ge \mu\Lambda_{\min}(R)\|K^l\|_2^2.\end{aligned}$$

This directly shows that $C(K^l) \to +\infty$ as $\|K^l\|_2 \to +\infty$. Next, we assume $K^l \to K \in \partial\mathbb{K}$. Based on Proposition 1, we know that for all l, there exists $Y^l \succ 0$ such that

$$Y^l - \mathcal{L}^{K^l}(Y^l) = Q + (K^l)^T R K^l,$$

where the dependence of $\mathcal{L}$ on K is emphasized by the superscript. We now want to show that the sequence $\{Y^l\}$ is unbounded, and will use a contradiction argument. Suppose that $\{Y^l\}$ is bounded. By the Weierstrass-Bolzano theorem, $\{Y^l\}$ admits a subsequence $\{Y^{l_n}\}_{n=0}^{\infty}$ which converges to some limit point denoted as Y.

Clearly, we have $Y \succeq 0$. For the same subsequence $\{l_n\}_{n=0}^{\infty}$, we have $\lim_{n\to\infty} K^{l_n} = K \in \partial\mathbb{K}$. For all l_n, we still have

$$Y^{l_n} - \mathcal{L}^{K^{l_n}}(Y^{l_n}) = Q + (K^{l_n})^T R K^{l_n}.$$

Now letting n go to ∞, by continuity, leads to the equation $Y - \mathcal{L}^K(Y) = Q + K^T R K$. Since $Q \succ 0$, $R \succ 0$, and $\mathcal{L}^K(Y) \succeq 0$, we conclude $Y \succ 0$. By Proposition 1, we have $K \in \mathbb{K}$, and this contradicts the fact that $K \in \partial\mathbb{K}$. Therefore, $\{Y^l\}$ must be unbounded. Since Y^l is positive definite, we can further conclude that $\{\operatorname{tr}\left(Y^l\right)\}$ is unbounded and $C(K^l) \to \infty$. This completes the proof of Statement 1.

Next, we prove Statement 2. Recall that we have $\Gamma_i = A_i - B_i K_i$. For simplicity, we denote $\Gamma_i' := A_i - B_i K_i'$. By definition, we have $C(K') - C(K) = \langle P^{K'} - P^K, X(0)\rangle$. Based on (10), we can show

$$\begin{aligned} P^{K'} - P^K =& \sum_{t=0}^{\infty} (\mathcal{L}^{K'})^t \Big(\Delta K^T (R + B^T \mathcal{E}(P^K) B)\Delta K \\ & + \Delta K^T L^K + (L^K)^T \Delta K\Big). \end{aligned}$$

Here the notation $\mathcal{L}^{K'}$ emphasizes that this is the operator associated with K'. Now we can prove Statement 2 by applying the above equation and the fact that $\mathcal{T}^{K'}$ is the adjoint operator of $\mathcal{L}^{K'}$.

To prove Statement 3, we rewrite the almost smoothness condition and complete the squares as follows

$$\begin{aligned} C(K') - C(K) =& -\langle (L^K)^T \Psi^{-1} L^K, \mathbf{X}^{K'}\rangle \\ & + \langle \left(-\Delta K + \Psi^{-1} L^K\right)^T \Psi \left(-\Delta K + \Psi^{-1} L^K\right), \mathbf{X}^{K'}\rangle, \end{aligned}$$

which leads to $C(K') - C(K) \geq -\langle (L^K)^T \Psi^{-1} L^K, \mathbf{X}^{K'}\rangle$. Then we can set $K' = K^*$ to prove Statement 3.

Statement 4 can be proved using the continuity and coercivity of $C(K)$. With the coercive property in place, we can continuously extend the function domain from $\mathbb{K}$ to $\mathbb{M}^{N_s}_{k\times d}$ by allowing ∞ as a function value. Based on Proposition 11.12 in [46], we know that $\mathbb{K}_\alpha$ is bounded for any finite α. Since $C(K)$ is continuous on $\mathbb{K}$, the set $\mathbb{K}_\alpha$ is also closed. Hence Statement 4 holds as desired.

Finally, to prove Statement 5, we only need to bound the norm of $\nabla^2 C(K)$. Then the desired conclusion follows by applying the mean value theorem. Since $\nabla^2 C(K)$ is self-adjoint, its operator norm can be characterized as

$$\|\nabla^2 C(K)\| = \sup_{\|E\|_2 = 1} |\nabla^2 C(K)[E, E]|.$$

Based on the Hessian formula (15), we have

$$\begin{aligned} &\sup_{\|E\|_2=1} |\nabla^2 C(K)[E,E]| \\ &\quad \leq 2 \sup_{\|E\|_2=1} |\langle (R + B^T \mathcal{E}(P^K) B) E \mathbf{X}^K, E\rangle| \\ &\quad\quad + 4 \sup_{\|E\|_2=1} |\langle B^T \mathcal{E}((P^K)')\Gamma \mathbf{X}^K, E\rangle|. \end{aligned} \tag{22}$$

Now we only need to provide upper bounds for the two terms on the right side of the above inequality. For simplicity, we denote $q_1 := \sup_{\|E\|_2=1} |\langle (R + B^T \mathcal{E}(P^K) B) E \mathbf{X}^K, E\rangle|$ and $q_2 := \sup_{\|E\|_2=1} |\langle B^T \mathcal{E}((P^K)')\Gamma \mathbf{X}^K, E\rangle|$. As a matter of fact, q_1 and q_2 can be bounded as follows

$$q_1 \leq \left(\|R\|_{\max} + \|B\|_{\max}^2 \frac{C(K)}{\mu}\right) \frac{C(K)}{\Lambda_{\min}(Q)} \tag{23}$$

$$q_2 \leq \frac{\xi \|B\|_{\max} C(K)^2}{\mu \Lambda_{\min}(Q)} \tag{24}$$

The proofs of (23) and (24) are tedious and hence are deferred to the appendix for readability. Now we are ready to prove the L-smoothness of $C(K)$ within the set $\mathbb{K}_\alpha$. Notice $C(K) \leq \alpha$ for any $K \in \mathbb{K}_\alpha$. Hence we can combine (23) and (24) to show $2q_1 + 4q_2 \leq L$ where L is given in Statement 5. Based on the mean value theorem, this leads to the desired conclusion. It is worth emphasizing that (21) only holds when the line segment between K and K' is in $\mathbb{K}_\alpha$. Since $\mathbb{K}_\alpha$ is non-convex in general, it is possible that there exists $K, K' \in \mathbb{K}_\alpha$ such that (21) does not hold. □

Now we briefly explain the importance of the above properties. When applying the gradient method to search for K^*, two issues need to be addressed and our techniques will heavily rely on the above cost properties.

1) Feasibility: One has to ensure that the iterates generated by the gradient method always stay in the non-convex feasible set $\mathbb{K}$. The coercivity implies that the function $C(K)$ serves as a barrier function on $\mathbb{K}$. Based on the coercivity and the compactness of the sublevel set, one can show that the decrease of the cost ensures the next iterate to stay inside $\mathbb{K}$.
2) Convergence: After ensuring the feasibility, one next needs to show that the iterates generated by the optimization method converge to K^*. The smoothness and gradient dominance properties will play a key role in the convergence proof when there is an absence of convexity.

IV. Gradient Method and Convergence

In Section II-C, we have reviewed policy optimization for the LTI case. In this section, we will consider the gradient method in the MJLS LQR setting and provide new global convergence guarantees. In the MJLS LQR setting, the gradient method iterates as

$$K^{n+1} = K^n - \eta \nabla C(K^n), \tag{25}$$

where K^0 is required to be in $\mathbb{K}$. The stepsize η is a hyperparameter to be tuned. When the parameters (A, B, Q, R) are exactly known, the gradient $\nabla C(K^n)$ can be evaluated using the formula (13). If the model parameters are unknown, one can still estimate the gradient from data using either zeroth-order optimization [41] or policy gradient theorem [47]. Now we present the convergence theory for the update rule (25) with exact gradient information. We first need to ensure that the iterates generated by (25) are always in $\mathbb{K}$. Consider the one-step gradient update $K' \leftarrow K - 2\eta L^K \mathbf{X}^K$. We will use the coercivity of $C(K)$ and the compactness of $\mathbb{K}_\alpha$ to show that for all $K \in \mathbb{K}$, we can choose η such that K' will also be in $\mathbb{K}$.

Lemma 3. *Suppose $K \in \mathbb{K}_\alpha$ and $K' = K - \eta \nabla C(K)$. Set L as described in Statement 5 of Lemma 2. If $0 < \eta \leq \frac{1}{L}$, then we have $K' \in \mathbb{K}_\alpha \subset \mathbb{K}$ and*

$$C(K') \leq C(K) - \frac{\eta}{2}\|\nabla C(K)\|_2^2. \tag{26}$$

Proof. We define the interior set of $\mathbb{K}_\alpha$ as $\mathbb{K}_\alpha^{\text{o}} := \{K \in \mathbb{K} : C(K) < \alpha\}$. The complement of $\mathbb{K}_\alpha^{\text{o}}$ is denoted as $(\mathbb{K}_\alpha^{\text{o}})^c$. Notice $\|\nabla^2 C(K)\| \leq L$ for all $K \in \mathbb{K}_\alpha$. By continuity, there exists $\epsilon > 0$ such that $\|\nabla^2 C(K)\| \leq 1.1L$ for all $K \in \mathbb{K}_{\alpha+\epsilon}$.

Clearly $(\mathbb{K}_{\alpha+\epsilon}^{\text{o}})^c$ is a closed set and $(\mathbb{K}_{\alpha+\epsilon}^{\text{o}})^c \cap \mathbb{K}_\alpha = \emptyset$. Since $\mathbb{K}_\alpha$ is compact, we know the distance between $\mathbb{K}_\alpha$ and $(\mathbb{K}_{\alpha+\epsilon}^{\text{o}})^c$ is strictly positive. We denote this distance as δ. Let us choose $\tau = \min\{0.9\delta/\|\nabla C(K)\|, 1/1.1L\}$. Obviously, the line segment between K and $(K - \tau\nabla C(K))$ is in $\mathbb{K}_{\alpha+\epsilon}$. Notice $\|\nabla^2 C(K)\| \leq 1.1L$ for all $K \in \mathbb{K}_{\alpha+\epsilon}$, and hence we have

$$\begin{aligned} C(K - \tau\nabla C(K)) \leq& C(K) + \langle \nabla C(K), K - \tau\nabla C(K) - K\rangle \\ & + \frac{1.1L}{2}\|K - \tau\nabla C(K) - K\|_2^2 \end{aligned}$$

which leads to

$$C(K-\tau\nabla C(K)) \le C(K) + \left(-\tau + \frac{1.1L\tau^2}{2}\right)\|\nabla C(K)\|_2^2$$

As long as $\tau \le 2/(1.1L)$, we have $-\tau + \frac{1.1L\tau^2}{2} \le 0$ and $C(K - \tau\nabla C(K)) \le C(K)$. Hence we have $K - \tau\nabla C(K) \in \mathbb{K}_\alpha$. Actually, it is straightforward to see that the line segment between K and $(K - \tau\nabla C(K))$ is in $\mathbb{K}_\alpha$.

The rest of the proof follows from induction. We can apply the same argument to show that the line segment between $(K - \tau\nabla C(K))$ and $(K - 2\tau\nabla C(K))$ is also in $\mathbb{K}_\alpha$. This means that the line segment between K and $(K - 2\tau\nabla C(K))$ is in $\mathbb{K}_\alpha$. Since $\tau > 0$, we only need to apply the above argument for finite times and then will be able to show that the line segment between K and $(K - \eta\nabla C(K))$ is in $\mathbb{K}_\alpha$ for any $0 < \eta \le \frac{1}{L}$.[2] Since $\|\nabla^2 C(K)\| \le L$ for all $K \in \mathbb{K}_\alpha$, we have

$$\begin{aligned}
C(K') &\le C(K) + \langle \nabla C(K), K' - K\rangle + \frac{L}{2}\|K' - K\|_2^2 \\
&= C(K) + \left(-\eta + \frac{L\eta^2}{2}\right)\|\nabla C(K)\|_2^2 \\
&\le C(K) - \frac{\eta}{2}\|\nabla C(K)\|_2^2,
\end{aligned}$$

where the last step follows from the fact that we have $0 < \eta \le \frac{1}{L}$. This completes the proof. □

Next, we can combine (26) with the gradient dominance property to show that the cost associated with the one-step progress of the gradient descent method is decreasing. This step is quite standard.

Lemma 4. *Suppose $K \in \mathbb{K}_\alpha$ and $K' = K - \eta\nabla C(K)$. Set L as described in Statement 5 of Lemma 2. If $0 < \eta \le \frac{1}{L}$, then the following inequality holds*

$$C(K') - C(K^*) \le \left(1 - \frac{2\mu^2\Lambda_{\min}(R)}{\|\mathbf{X}^{K^*}\|_{\max}}\eta\right)(C(K) - C(K^*)).$$

Proof. By Lemma 3, we know K' is stabilizing. We can combine (26) with Statement 3 in Lemma 2 to show

$$\begin{aligned}
C(K') - C(K) &\le -\frac{\eta}{2}\|\nabla C(K)\|_2^2 \\
&\le -\frac{2\mu^2\Lambda_{\min}(R)\eta}{\|\mathbf{X}^{K^*}\|_{\max}}\left(C(K) - C(K^*)\right)
\end{aligned}$$

which directly leads to the desired conclusion. □

Now we are ready to prove the global convergence of the policy gradient method (25).

Theorem 1. *Suppose $K^0 \in \mathbb{K}$. Choose $\alpha = C(K^0)$ and set L as described in Statement 5 of Lemma 2. For any step size $0 < \eta \le \frac{1}{L}$, the iterations generated by the gradient descent method (25) always stay in $\mathbb{K}$ and converge to the global minimum K^* linearly as follows*

$$\begin{aligned}
C(K^n) - C(K^*) \le \left(1 - \frac{2\mu^2\Lambda_{\min}(R)}{\|\mathbf{X}^{K^*}\|_{\max}}\eta\right)^n \times \\
\left(C(K^0) - C(K^*)\right). \qquad (27)
\end{aligned}$$

Proof. We will use an induction argument. Since $\alpha = C(K^0)$, we have $K^0 \in \mathbb{K}_\alpha$. By Lemma 4, we know (27) holds for $n = 1$. Since $C(K^1) \le C(K^0)$, we have $K^1 \in \mathbb{K}_\alpha$. We can apply Lemma 4 again to show (27) holds for $n = 2$. Now it is clear that we can repeatedly apply the above argument to show (27) holds for any n. □

From the above proof, one can see that without using projection, one can still guarantee the gradient method will stay in the feasible set and converge to the global minimum. When the model parameters are known, there are many other methods which can be used to solve K^* [27], [48]. We do not claim that the gradient method is more desirable than other methods when the model information is known. The purpose of our study is to bring new insights for understanding policy-based RL methods in the MJLS setting. When the model is unknown, one can still apply model-free techniques such as zeroth-order optimization [49], [50] to estimate the gradient $\nabla C(K)$ from data. Then the gradient estimation errors have to be explicitly addressed, and this has recently been done in a follow-up work [41]. The analysis in [41] combines the theory in our paper with some estimation error bounds to handle the model-free case.

Remark 1. The above proof technique is more general than the implicit regularization arguments in our previous conference paper [40] which addresses the convergence of the natural gradient method[3]which iterated as $K^{n+1} = K^n - \eta\nabla C(K^n)(\mathbf{X}^{K^n})^{-1}$. In [40], it has been shown that the natural gradient method with any $0 < \eta \le \frac{1}{2}\left(\|R\|_{\max} + \frac{\|B\|_{\max}^2 C(K^0)}{\mu}\right)^{-1}$ always stays in $\mathbb{K}$ and will converge to the global minimum K^* linearly as follows

$$C(K^n) - C(K^*) \le \left(1 - \frac{2\mu\Lambda_{\min}(R)}{\|\mathbf{X}^{K^*}\|_{\max}}\eta\right)^n (C(K^0) - C(K^*)).$$

The proof in [40] relies on an implicit regularization argument where P^K is used to construct a Lyapunov function for K' and then guarantee $K' \in \mathbb{K}$. Similar ideas have been used to show the convergence properties of policy optimization methods for the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ state feedback design problem [23]. However, such an implicit regularization proof idea does not work for the gradient method. For the gradient method, the value function at step n cannot be directly used as a Lyapunov function at step $(n+1)$. A similar fact has also been observed for the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control problem [23].

References

[1] R. Sutton and A. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[2] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, “High-dimensional continuous control using generalized advantage estimation,” in *International Conference on Learning Representation*, 2015.

[3] S. Levine, C. Finn, T. Darrell, and P. Abbeel, “End-to-end training of deep visuomotor policies,” *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1334–1373, 2016.

[4] Y. Duan, X. Chen, R. Houthooft, J. Schulman, and P. Abbeel, “Benchmarking deep reinforcement learning for continuous control,” in *International Conference on Machine Learning*, 2016, pp. 1329–1338.

[5] R. Sutton, D. McAllester, S. Singh, and Y. Mansour, “Policy gradient methods for reinforcement learning with function approximation,” in *Advances in neural information processing systems*, 2000, pp. 1057–1063.

[6] S. Kakade, “A natural policy gradient,” in *Advances in neural information processing systems*, 2002, pp. 1531–1538.

[7] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, “Trust region policy optimization,” in *International Conference on Machine Learning*, 2015, pp. 1889–1897.

[8] J. Peters and S. Schaal, “Natural actor-critic,” *Neurocomputing*, vol. 71, no. 7-9, pp. 1180–1190, 2008.

[9] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” *arXiv preprint arXiv:1707.06347*, 2017.

[10] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, “Deep reinforcement learning that matters,” in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[2] The argument even works for any $\eta \le \frac{2}{1.1L}$. Since the step size leading to the fastest convergence rate is $\frac{1}{L}$, we state our result only for $\eta \le \frac{1}{L}$.

[3] See Appendix A.1 in the `arXiv` version of [47] for more explanation of the terminology “natural gradient.”

[11] A. Rajeswaran, K. Lowrey, E. Todorov, and S. Kakade, "Towards generalization and simplicity in continuous control," in *Advances in Neural Information Processing Systems*, 2017, pp. 6550–6561.

[12] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, 2018, pp. 1467–1476.

[13] J. Bu, A. Mesbahi, M. Fazel, and M. Mesbahi, "LQR through the lens of first order methods: Discrete-time case," *arXiv preprint arXiv:1907.08921*, 2019.

[14] D. Malik, A. Pananjady, K. Bhatia, K. Khamaru, P. Bartlett, and M. Wainwright, "Derivative-free methods for policy optimization: Guarantees for linear quadratic systems," *arXiv preprint arXiv:1812.08305*, 2018.

[15] S. Tu and B. Recht, "The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint," *arXiv preprint arXiv:1812.03565*, 2018.

[16] Z. Yang, Y. Chen, M. Hong, and Z. Wang, "On the global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost," *arXiv preprint arXiv:1907.06246*, 2019.

[17] K. Krauth, S. Tu, and B. Recht, "Finite-time analysis of approximate policy iteration for the linear quadratic regulator," in *Advances in Neural Information Processing Systems*, 2019, pp. 8512–8522.

[18] H. Mohammadi, A. Zare, M. Soltanolkotabi, and M. Jovanović, "Convergence and sample complexity of gradient methods for the model-free linear quadratic regulator problem," *arXiv preprint arXiv:1912.11899*, 2019.

[19] H. Mohammadi, A. Zare, M. Soltanolkotabi, and M. Jovanovic, "Global exponential convergence of gradient methods over the nonconvex landscape of the linear quadratic regulator," in *2019 IEEE 58th Conference on Decision and Control (CDC)*, 2019.

[20] I. Fatkhullin and B. Polyak, "Optimizing static linear feedback: Gradient method," *SIAM Journal on Control and Optimization*, vol. 59, no. 5, pp. 3887–3911, 2021.

[21] L. Furieri, Y. Zheng, and M. Kamgarpour, "Learning the globally optimal distributed LQ regulator," in *Learning for Dynamics and Control*, 2020, pp. 287–297.

[22] Y. Li, Y. Tang, R. Zhang, and N. Li, "Distributed reinforcement learning for decentralized linear quadratic control: A derivative-free policy optimization approach," *arXiv preprint arXiv:1912.09135*, 2019.

[23] K. Zhang, B. Hu, and T. Başar, "Policy optimization for $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$ robustness guarantee: Implicit regularization and global convergence," *arXiv preprint arXiv:1910.09496*, 2019.

[24] B. Gravell, P. M. Esfahani, and T. Summers, "Learning optimal controllers for linear systems with multiplicative noise via policy gradient," *IEEE Transactions on Automatic Control*, vol. 66, no. 11, pp. 5283–5298, 2021.

[25] K. Zhang, B. Hu, and T. Basar, "On the stability and convergence of robust adversarial reinforcement learning: A case study on linear quadratic systems," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[26] G. Qu, C. Yu, S. Low, and A. Wierman, "Combining model-based and model-free methods for nonlinear control: A provably convergent policy gradient approach," *arXiv preprint arXiv:2006.07476*, 2020.

[27] O. Costa, M. Fragoso, and R. Marques, *Discrete-time Markov jump linear systems*. Springer London, 2006.

[28] Y. Bar-Shalom and X. Li, "Estimation and tracking- principles, techniques, and software," *Norwood, MA: Artech House, Inc, 1993.*, 1993.

[29] E. Fox, E. S. M. Jordan, and A. Willsky, "Bayesian nonparametric inference of switching dynamic linear models," *IEEE Transactions on Signal Processing*, vol. 59, no. 4, pp. 1569 – 1585, 2011.

[30] K. Gopalakrishnan, H. Balakrishnan, and R. Jordan, "Stability of networked systems with switching topologies," in *IEEE Conference on Decision and Control*, 2016, pp. 1889–1897.

[31] V. Pavlovic, J. Rehg, and J. MacCormick, "Learning switching linear models of human motion," in *Advances in Neural Information Processing Systems*, 2000.

[32] D. Sworder and J. Boyd, *Estimation problems in hybrid systems*. Cambridge University Press, 1999.

[33] A. N. Vargas, E. F. Costa, and J. B. R. do Val, "On the control of Markov jump linear systems with no mode observation: Application to a DC motor device," *International Journal of Robust and Nonlinear Control*, vol. 23, no. 10, pp. 1136–1150, 2013.

[34] B. Hu, P. Seiler, and A. Rantzer, "A unified analysis of stochastic optimization methods using jump system theory and quadratic constraints," in *Conference on Learning Theory*, 2017, pp. 1157–1189.

[35] B. Hu and U. Syed, "Characterizing the exact behaviors of temporal difference learning algorithms using Markov jump linear system theory," in *Advances in Neural Information Processing Systems*, 2019, pp. 8477–8488.

[36] O. L. Costa and J. C. Aya, "Monte Carlo TD (λ)-methods for the optimal control of discrete-time Markovian jump linear systems," *Automatica*, vol. 38, no. 2, pp. 217–225, 2002.

[37] R. L. Beirigo, M. G. Todorov, and A. d. M. S. Barreto, "Online TD (λ) for discrete-time Markov jump linear systems," in *2018 IEEE Conference on Decision and Control (CDC)*, 2018, pp. 2229–2234.

[38] M. Schuurmans, P. Sopasakis, and P. Patrinos, "Safe learning-based control of stochastic jump linear systems: a distributionally robust approach," in *2019 IEEE 58th Conference on Decision and Control (CDC)*, 2019, pp. 6498–6503.

[39] A. N. Vargas, D. C. Bortolin, E. F. Costa, and J. B. do Val, "Gradient-based optimization techniques for the design of static controllers for Markov jump linear systems with unobservable modes," *International Journal of Numerical Modelling: Electronic Networks, Devices and Fields*, vol. 28, no. 3, pp. 239–253, 2015.

[40] J. P. Jansch-Porto, B. Hu, and G. E. Dullerud, "Convergence guarantees of policy optimization methods for Markovian jump linear systems," in *2020 American Control Conference (ACC)*, 2020, pp. 2882–2887.

[41] S. Rathod, M. Bhadu, and A. De, "Global convergence using policy gradient methods for model-free Markovian jump linear quadratic control," *arXiv preprint arXiv:2111.15228*, 2021.

[42] R. Abraham, J. E. Marsden, and T. Ratiu, *Manifolds, tensor analysis, and applications*. Springer Science & Business Media, 2012, vol. 75.

[43] M. Fragoso, "Discrete-time jump LQG problem," *International Journal of Systems Science*, vol. 20, no. 12, pp. 2539–2545, 1989.

[44] K. Mårtensson and A. Rantzer, "Gradient methods for iterative distributed control synthesis," in *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*, 2009, pp. 549–554.

[45] J. Dattorro, *Convex optimization & Euclidean distance geometry*. Lulu. com, 2010.

[46] H. H. Bauschke, P. L. Combettes, *et al.*, *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, 2011, vol. 408.

[47] J. P. Jansch-Porto, B. Hu, and G. Dullerud, "Policy learning of MDPs with mixed continuous/discrete variables: A case study on model-free control of Markovian jump systems," in *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, 2020, pp. 947–957.

[48] M. A. Rami and L. E. Ghaoui, "LMI optimization for nonstandard riccati equations arising in stochastic control," *IEEE transactions on automatic control*, vol. 41, no. 11, pp. 1666–1671, 1996.

[49] A. Conn, K. Scheinberg, and L. Vicente, *Introduction to derivative-free optimization*. Siam, 2009, vol. 8.

[50] Y. Nesterov and V. Spokoiny, "Random gradient-free minimization of convex functions," *Foundations of Computational Mathematics*, vol. 17, no. 2, pp. 527–566, 2017.

Appendix

A. Proof of the Bound (23)

The proof of (23) is straightforward. Notice that we have

$$\begin{aligned}
&\langle (R + B^T \mathcal{E}(P^K)B) E \mathbf{X}^K,\, E\rangle \\
&\quad = \sum_{i\in\Omega} \operatorname{tr}\Big(E_i^T (R_i + B_i^T \mathcal{E}_i(P^K) B_i) E_i \mathbf{X}_i^K \Big) \\
&\quad \le \sum_{i\in\Omega} \| E_i^T (R_i + B_i^T \mathcal{E}_i(P^K) B_i) E_i \| \operatorname{tr}\Big(\mathbf{X}_i^K\Big) \\
&\quad \le \sum_{i\in\Omega} \|E_i\|^2 \| R_i + B_i^T \mathcal{E}_i(P^K) B_i \| \operatorname{tr}\Big(\mathbf{X}_i^K\Big) \\
&\quad \le \|E\|_{\max}^2 \| R + B^T \mathcal{E}(P^K) B\|_{\max} \sum_{i\in\Omega} \operatorname{tr}\Big(\mathbf{X}_i^K\Big) \\
&\quad \le \|E\|_2^2 \Big(\|R\|_{\max} + \|B\|_{\max}^2 \|P^K\|_{\max} \Big) \sum_{i\in\Omega} \operatorname{tr}\Big(\mathbf{X}_i^K\Big)
\end{aligned}$$

where the last step follows from (11). Hence we immediately have

$$q_1 \le \Big(\|R\|_{\max} + \|B\|_{\max}^2 \|P^K\|_{\max} \Big) \sum_{i\in\Omega} \operatorname{tr}\Big(\mathbf{X}_i^K\Big). \qquad \text{(A.1)}$$

Now what we need to bound $\|P^K\|_{\max}$ and $\sum_{i\in\Omega}\operatorname{tr}\left(\mathbf{X}_i^K\right)$. Recall $C(K)=\mathbb{E}_{x_0\sim\mathcal{D}}\left[\operatorname{tr}\left(\left(\sum_{i\in\Omega}\pi_iP_i^K\right)x_0x_0^T\right)\right]$. Therefore, we have

$$\begin{aligned}C(K)&\geq\operatorname{tr}\left(\sum_{i\in\Omega}\pi_iP_i^K\right)\sigma_{\min}\left(\mathbb{E}_{x_0\sim\mathcal{D}}\left[x_0x_0^T\right]\right)\\&\geq\left(\sum_{i\in\Omega}\operatorname{tr}\left(P_i^K\right)\right)\min_{i\in\Omega}(\pi_i)\sigma_{\min}\left(\mathbb{E}_{x_0\sim\mathcal{D}}\left[x_0x_0^T\right]\right)\end{aligned}$$

which leads to the following upper bound

$$\|P^K\|_{\max}\leq\sum_{i\in\Omega}\operatorname{tr}\left(P_i^K\right)\leq\frac{C(K)}{\mu}.\tag{A.2}$$

Notice $\mathcal{T}$ is the adjoint operator of $\mathcal{L}$. Hence we also have

$$\begin{aligned}C(K)=\langle P^K,\,X(0)\rangle&=\langle\sum_{t=0}^{\infty}\mathcal{L}^t(Q+K^TRK),\,X(0)\rangle\\&=\langle Q+K^TRK,\,\sum_{t=0}^{\infty}\mathcal{T}^t(X(0))\rangle\\&=\langle Q+K^TRK,\,\mathbf{X}^K\rangle\\&\geq\sum_{i\in\Omega}\sigma_{\min}(Q_i)\operatorname{tr}\left(\mathbf{X}_i^K\right)\\&\geq\Lambda_{\min}(Q)\sum_{i\in\Omega}\operatorname{tr}\left(\mathbf{X}_i^K\right),\end{aligned}$$

which leads to another useful bound

$$\sum_{i\in\Omega}\operatorname{tr}\left(\mathbf{X}_i^K\right)\leq\frac{C(K)}{\Lambda_{\min}(Q)}.\tag{A.3}$$

Substituting (A.2) and (A.3) into (A.1) leads to (23). □

B. Proof of the Bound (24)

For simplicity, we shorten the notation $(P^K)'[E]$ as $(P^K)'$. To prove (24), first notice that we can use the Cauchy-Schwarz inequality to show

$$\begin{aligned}&\langle B^T\mathcal{E}((P^K)')\varGamma\mathbf{X}^K,\,E\rangle\\&\quad=\langle E^TB^T\mathcal{E}((P^K)')\varGamma(\mathbf{X}^K)^{1/2},\,(\mathbf{X}^K)^{1/2}\rangle\\&\quad\leq\|E^TB^T\mathcal{E}((P^K)')\varGamma(\mathbf{X}^K)^{1/2}\|_2\|(\mathbf{X}^K)^{1/2}\|_2.\end{aligned}$$

Next, we bound $\|E^TB^T\mathcal{E}((P^K)')\varGamma(\mathbf{X}^K)^{1/2}\|_2$ as follows

$$\begin{aligned}&\|E^TB^T\mathcal{E}((P^K)')\varGamma(\mathbf{X}^K)^{1/2}\|_2^2\\&=\sum_{i\in\Omega}\operatorname{tr}\left(\mathcal{E}_i((P^K)')B_iE_iE_i^TB_i^T\mathcal{E}_i((P^K)')\varGamma_i\mathbf{X}_i^K\varGamma_i^T\right)\\&=\sum_{i\in\Omega}\|B_i\|^2\|E_i\|^2\operatorname{tr}\left(\mathcal{E}_i((P^K)')\mathcal{E}_i((P^K)')\varGamma_i\mathbf{X}_i^K\varGamma_i^T\right)\\&\leq\|B\|_{\max}^2\|E\|_{\max}^2\sum_{i\in\Omega}\operatorname{tr}\left(\mathcal{E}_i((P^K)')\mathcal{E}_i((P^K)')\varGamma_i\mathbf{X}_i^K\varGamma_i^T\right)\\&\leq\|B\|_{\max}^2\|E\|_2^2\sum_{i\in\Omega}\operatorname{tr}\left(\mathcal{E}_i((P^K)')\mathcal{E}_i((P^K)')\varGamma_i\mathbf{X}_i^K\varGamma_i^T\right)\end{aligned}\tag{B.1}$$

Since $\mathcal{E}_i((P^K)')\mathcal{E}_i((P^K)')$ is positive semidefinite, we have

$$\begin{aligned}&\sum_{i\in\Omega}\operatorname{tr}\left(\mathcal{E}_i((P^K)')\mathcal{E}_i((P^K)')\varGamma_i\mathbf{X}_i^K\varGamma_i^T\right)\\&\qquad\leq\sum_{i\in\Omega}\|\mathcal{E}_i((P^K)')\|^2\operatorname{tr}\left(\varGamma_i\mathbf{X}_i^K\varGamma_i^T\right)\\&\qquad\leq\|(P^K)'\|_{\max}^2\sum_{i\in\Omega}\operatorname{tr}\left(\varGamma_i\mathbf{X}_i^K\varGamma_i^T\right)\end{aligned}$$

If $K\in\mathbb{K}$, we know $\mathcal{T}(\mathbf{X}^K)-\mathbf{X}^K\prec 0$ and hence the following also holds

$$\begin{aligned}\sum_{i\in\Omega}\operatorname{tr}\left(\varGamma_i\mathbf{X}_i^K\varGamma_i^T\right)&=\sum_{j\in\Omega}\operatorname{tr}\left(\sum_{i\in\Omega}p_{ij}\varGamma_i\mathbf{X}_i^K\varGamma_i^T\right)\\&\leq\sum_{j\in\Omega}\operatorname{tr}\left(\mathcal{T}_j\left(\mathbf{X}^K\right)\right)\leq\sum_{j\in\Omega}\operatorname{tr}\left(\mathbf{X}_j^K\right)\end{aligned}$$

Therefore, substituting the above bounds into (B.1) leads to

$$\begin{aligned}&\|E^TB^T\mathcal{E}((P^K)')\varGamma(\mathbf{X}^K)^{1/2}\|_2^2\\&\qquad\leq\|B\|_{\max}^2\|E\|_2^2\|(P^K)'\|_{\max}^2\sum_{j\in\Omega}\operatorname{tr}\left(\mathbf{X}_j^K\right)\end{aligned}$$

Since $\|(\mathbf{X}^K)^{1/2}\|_2^2=\sum_{j\in\Omega}\operatorname{tr}\left(\mathbf{X}_j^K\right)$, we finally have

$$\begin{aligned}&\langle B^T\mathcal{E}((P^K)')\varGamma\mathbf{X}^K,\,E\rangle\\&\qquad\leq\|B\|_{\max}\|E\|_2\|(P^K)'\|_{\max}\sum_{j\in\Omega}\operatorname{tr}\left(\mathbf{X}_j^K\right)\end{aligned}\tag{B.2}$$

Based on (B.2), proving (24) only requires showing that the following bound holds for any $\|E\|_2=1$ and $K\in\mathbb{K}$,

$$\|(P^K)'[E]\|_{\max}\leq\xi\|P^K\|_{\max},\tag{B.3}$$

where ξ is given as

$$\xi=\frac{1}{\Lambda_{\min}(Q)}\left(\frac{1+\|B\|_{\max}^2}{\mu}C(K)+\|R\|_{\max}\right)-1.$$

Once (B.3) is proved, it can be combined with (B.2), (A.2), and (A.3) to verify (24) easily.

Now the only remaining task is to prove (B.3). Let us first show $(P^K)'[E]\leq\xi P^K$ given $\|E\|_2=1$. We will use Corollary 2.7 in [27] which states that $\tilde{X}\succeq X$ if $(X,\tilde{X})$ satisfy $X-\mathcal{L}(X)=S$ and $\tilde{X}-\mathcal{L}(\tilde{X})=\tilde{S}$ with $\tilde{S}\succeq S$ and $K\in\mathbb{K}$. Since $\mathcal{E}(P^K)\succ 0$ and $R\succ 0$, we have

$$\begin{aligned}&(P^K)'[E]-\mathcal{L}((P^K)'[E])\\&=(-BE)^T\mathcal{E}(P^K)\varGamma+\varGamma^T\mathcal{E}(P^K)(-BE)+E^TRK+K^TRE\\&\preceq\mathcal{L}(P^K)+(BE)^T\mathcal{E}(P^K)BE+K^TRK+E^TRE\\&=P^K-Q+(BE)^T\mathcal{E}(P^K)BE+E^TRE\\&=:W\end{aligned}$$

If we can show $W\preceq\xi(Q+K^TRK)$, then Corollar 2.7 in [27] can be directly applied to show $(P^K)'[E]\leq\xi P^K$. Note that we have the following upper bound,

$$\begin{aligned}&\|P_i^K+E_i^TB_i^T\mathcal{E}_i(P^K)B_iE_i\|\\&\qquad\leq\|P_i^K\|+\|E_i^TB_i^T\mathcal{E}_i(P^K)B_iE_i\|\\&\qquad\leq\|P^K\|_{\max}+\|E\|_{\max}^2\|B\|_{\max}^2\|P^K\|_{\max}\\&\qquad\leq(1+\|B\|_{\max}^2)\frac{C(K)}{\mu}\end{aligned}$$

which directly leads to the following result

$$\begin{aligned}W&\preceq\left((1+\|B\|_{\max}^2)\frac{C(K)}{\mu}+\|R\|_{\max}\right)\mathcal{I}-Q\\&\preceq\frac{1}{\Lambda_{\min}(Q)}\left((1+\|B\|_{\max}^2)\frac{C(K)}{\mu}+\|R\|_{\max}\right)Q-Q\\&=\xi Q\end{aligned}$$

where $\mathcal{I}:=(I,\ldots,I)\in\mathbb{M}_{d\times d}^{N_s}$ Notice the bound makes sense since we know $C(K)\geq\Lambda_{\min}(Q)\mu$. Therefore, we have $(P^K)'[E]\preceq\xi P^K$. This directly leads to (B.3). Now we can complete the proof by combining (B.3), (B.2), (A.2), and (A.3). □