2024 American Control Conference (ACC)
July 8-12, 2024. Toronto, Canada

A Case Study on the Convergence of Direct Policy Search for Linear Quadratic Gaussian Control

Darioush Keivan[1], Peter Seiler[2], Geir Dullerud[1], and Bin Hu[3]

Abstract— **Policy optimization has gained renewed attention from the control community, serving as a pivotal link between control theory and reinforcement learning. In the past few years, the global convergence theory of direct policy search on state-feedback linear control benchmarks has been developed. However, it remains difficult to establish the global convergence of policy optimization on the linear quadratic Gaussian (LQG) problem, marked by the presence of suboptimal stationary points and the lack of cost coerciveness. In this paper, we revisit the policy optimization intricacies of LQG via a case study on first-order single-input single-output (SISO) systems. For this case study, while the issue related to suboptimal stationary points can be easily fixed via parameterizing the policy class more carefully, the non-coerciveness of the LQG cost function still poses a substantial obstacle to a straightforward global convergence proof for the policy gradient method. Our contribution, within the scope of this case study, introduces an approach to construct a positive invariant set for the policy gradient flow, addressing the non-coerciveness issue in the global convergence proof. Based on our analysis, the policy gradient flow can be guaranteed to converge to the globally optimal full-order dynamic controller in this particular scenario. In summary, although centered on a specific case study, our work broadens the comprehension of how the absence of coerciveness impacts LQG policy optimization, highlighting inherent complexities.**

I. INTRODUCTION

The empirical successes of deep reinforcement learning [1], [2] have sparked considerable interest in direct policy search techniques within the field of control [3]. Substantial progress has been made in comprehending the global convergence properties of direct policy search across various linear state-feedback control problems such as the linear quadratic regulator (LQR) [4]–[10], risk-sensitive control [11]–[15], linear quadratic dynamic games [16]–[18], Markov jump linear control [19]–[21], stabilization [22]–[24], and nonsmooth $\mathcal{H}_\infty$ synthesis [25]. More recently, the research attention has been shifted towards the output feedback setting [26]–[32], with a particular emphasis on the linear quadratic Gaussian (LQG) problem [33]–[36]. For the LQG problem, globally optimal controllers emerge as dynamical controllers, deduced by solving two algebraic Riccati equations [37]. Although the LQG problem has been extensively studied in classical control, its optimization landscape over the policy space is less explored. Recent studies [33], [34] delved into the connectivity of the feasible set of full-order stabilizing dynamical controllers, revealing that while potentially disconnected, they encompass at most two path-connected components that are diffeomorphic under a similarity transformation, holding promise for gradient-based algorithms in finding a global optimal policy. However, it remains difficult to establish the global convergence of policy optimization for the LQG problem, marked by two important issues: the presence of suboptimal stationary points and the lack of cost coerciveness.

To better understand the theoretical properties of direct policy search on LQG, this paper presents a case study of a specific single-input single-output (SISO) system, which was originally studied as Example 3 of [33]. This example was previously utilized to demonstrate the existence of suboptimal stationary points in the LQG policy optimization problem[1]. For the system in our case study, we show that the issue related to suboptimal stationary points can be easily fixed via adopting a controllable policy parameterization. However, the non-coerciveness of the LQG cost function still poses a substantial obstacle to proving the global convergence of direct policy search on such a seemingly simple example. Our contribution, within the scope of this case study, introduces an approach to construct a positive invariant set for the policy gradient flow, addressing the non-coerciveness issue in the global convergence proof. Based on our analysis, the policy gradient flow can be guaranteed to converge to the globally optimal full-order dynamic controller in this particular scenario. Our result sheds new light on how to address the non-coerciveness issue in the LQG policy optimization problem.

II. BACKGROUND AND PROBLEM FORMULATION

A. Background: Policy Optimization for LQG

In this section, we briefly review the LQG policy optimization formulation from [33]. Consider a continuous-time linear dynamical system

$$\begin{aligned}\dot{x}(t) &= A\,x(t) + B\,u(t) + w(t)\\ y(t) &= C\,x(t) + v(t)\end{aligned} \tag{1}$$

where $x(t) \in \mathbb{R}^n$ is the system state, $u(t) \in \mathbb{R}^m$ is the control input, $y(t) \in \mathbb{R}^p$ is the output measurement, $w(t) \in \mathbb{R}^n$ is the system process noise, and $v(t) \in \mathbb{R}^p$ is the

[1] Darioush Keivan and Geir Dullerud are with the Coordinated Science Laboratory (CSL) and the Department of Mechanical Science & Engineering, University of Illinois at Urbana-Champaign, {dk12, dullerud}@illinois.edu

[2] Peter Seiler is with the Department of Electrical Engineering and Computer Science, University of Michigan, pseiler@umich.edu

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

[1] As discussed in [33], [35], suboptimal stationary points are tied to a loss of minimality (controllability or observability). Indeed, full-order minimal controllers cannot be saddle points [33].

979-8-3503-8265-5/$31.00 ©2024 AACC

measurement noise. It is assumed that both $w(t)$ and $v(t)$ are white Gaussian noises with intensity matrices $W \succeq 0$ and $V \succ 0$. The LQG cost is defined as

$$J := \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\int_0^T \left(x^\top Qx + u^\top Ru\right)dt\right] \tag{2}$$

where $Q \succeq 0$ and $R \succ 0$. The standard LQG problem assumes controllability for the pairs (A, B) and $(A, W^{\frac{1}{2}})$, and observability for the pairs (C, A) and $(Q^{\frac{1}{2}}, A)$. Then LQG can be tackled by solving two algebraic Riccati equations [37], yielding the following optimal controller:

$$\begin{aligned} \dot{\xi}(t) &= (A - BK - LC)\,\xi(t) + L\,y(t), \\ u(t) &= -K\,\xi(t), \end{aligned} \tag{3}$$

where $\xi(t) \in \mathbb{R}^n$ is the internal state of the controller, $K \in \mathbb{R}^{m\times n}$ is the feedback gain, and $L \in \mathbb{R}^{n\times p}$ is the Kalman gain. It is well-known that K and L can be evaluated as $K = R^{-1}B^TS$ and $L = PC^TV^{-1}$, respectively, with S and P being the unique positive semi-definite solutions to the following Riccati equations:

$$SA + A^TS - SBR^{-1}B^TS + Q = 0, \tag{4}$$

$$AP + PA^T - PC^TV^{-1}CP + W = 0. \tag{5}$$

Drawing from (3), we can formulate the LQG policy optimization problem by focusing solely on full-order dynamical controllers that are parameterized as

$$\begin{aligned} \dot{\xi}(t) &= A_K\,\xi(t) + B_K\,y(t), \\ u(t) &= C_K\,\xi(t), \end{aligned} \tag{6}$$

where $A_K \in \mathbb{R}^{n\times n}$, $B_K \in \mathbb{R}^{n\times p}$, and $C_K \in \mathbb{R}^{m\times n}$ are the decision variables to be solved. Then, LQG can be formulated as the following policy optimization problem

$$\min_{K\in\mathcal{K}} J(K) \tag{7}$$

where $K := (A_K, B_K, C_K)$ is any full-order dynamical controller parameterized by (6), J is the LQG cost function, and $\mathcal{K}$ is the set of all stabilizing full-order dynamical policies, i.e. $\mathcal{K} = \{(A_K, B_K, C_K) : A_{cl,K} \text{ is Hurwitz}\}$ with $A_{cl,K}$ being defined as

$$A_{cl,K} := \begin{bmatrix} A & BC_K \\ B_KC & A_K \end{bmatrix}.$$

For any stabilizing policy $K \in \mathcal{K}$, the cost J is given by

$$J(K) = \text{tr}(Q_{cl,K}X_K) = \text{tr}(W_{cl,K}Y_K) \tag{8}$$

where X_K and Y_K are the unique positive semi-definite solutions to the following Lyapunov equations

$$\begin{aligned} A_{cl}X_K + X_KA_{cl}^T + W_{cl,K} = 0 \\ A_{cl}^TY_K + Y_KA_{cl} + Q_{cl,K} = 0. \end{aligned} \tag{9}$$

In the above equations, $Q_{cl,K}$ and $W_{cl,K}$ are defined as

$$Q_{cl,K} := \begin{bmatrix} Q & 0 \\ 0 & C_K^TWC_K \end{bmatrix}, W_{cl,K} := \begin{bmatrix} W & 0 \\ 0 & B_KVB_K^T \end{bmatrix}.$$

For the policy search problem (7), there are two main issues. First, there may exist suboptimal stationary points that prevent convergence to the global minimum. Second, the LQG cost can be non-coercive [3], [33], and it is not even clear how to guarantee that direct policy search can stay in the feasible set $\mathcal{K}$. These challenges motivate our study, aiming for a better understanding.

B. Problem Statement: Case Study on an Example from [33]

In this paper, we will re-examine the above difficulties via a case study on a specific SISO system that was originally presented as Example 3 in [33]. In this particular example, we have $A = -1$, $B = 1$, $C = 1$, $W = V = 1$, and further assume $Q = 1$ and $R = 1$. In this setting, we can easily address the existence of suboptimal stationary points[2] via choosing $B_K = 1$. Such a simplified policy parameterization is actually natural for this example, since there is some redundancy in using both parameters (B_K, C_K) due to the fact that the transfer function of the controller is just equal to $K(s) = (B_KC_K)(sI - A_K)^{-1}$. Using $B_K = 1$, the feasibility set $\mathcal{K}$ in our problem can be simplified as

$$\mathcal{K} = \left\{\begin{bmatrix} A_K \\ C_K \end{bmatrix} \in \mathbb{R}^2 : A_K < 1, A_K + C_K < 0\right\}. \tag{10}$$

For simplicity, we will denote A_K as scalar a and C_K as scalar c, respectively. Within the feasible set (10), the LQG cost (8) can be calculated as

$$J = -\frac{-a^2 + ac^2 + a + c^3 - 3c^2 + c}{2(a + c)(a - 1)} \tag{11}$$

The derivatives of J with respect to a and c are given by

$$\begin{aligned} \frac{\partial J}{\partial a} &= \frac{c(a^2c + a^2 + 2ac^2 - 6ac + c^3 - 3c^2 + 4c)}{2(a + c)^2(a - 1)^2}, \\ \frac{\partial J}{\partial c} &= -\frac{2a^2c + a^2 + 4ac^2 - 6ac + 2c^3 - 3c^2}{2(a + c)^2(a - 1)}. \end{aligned} \tag{12}$$

We will see that even for this seemingly simple example, the lack of coerciveness still causes significant trouble for proving a global convergence result. We believe that the insights gained from this example will advance our understanding of the non-coercive issue in LQG policy optimization.

III. Optimization Landscape

In this section, we explore the optimization landscape of the above SISO example in more detail. From [33], it is known that the feasible set (10) is connected, and the LQG cost (11) is non-coercive over the feasible set. The following lemma provides a characterization of the stationary points for the above example, revealing the subtle optimization landscape caused by the non-coerciveness issue.

Lemma 1 (Non-coerciveness): Within the set (10), the cost (11) has a single stationary point (which is the global optimal point). However, at the boundary of the feasible set (10), there exists a point where the cost function retains a finite value, yet the corresponding gradient is indefinable.

[2] Based on [33], any point satisfying $B_K = C_K = 0$ and $A_K < 0$ is a suboptimal stationary point.

Proof: Assume $a+c\neq 0$ and $a\neq 1$ so that ∇J is well defined. Setting $\frac{\partial J}{\partial a}=0$ and $\frac{\partial J}{\partial c}=0$, we have

$$\begin{bmatrix} c(a^2c+a^2+2ac^2-6ac+c^3-3c^2+4c) \\ 2a^2c+a^2+4ac^2-6ac+2c^3-3c^2 \end{bmatrix}=\begin{bmatrix}0\\0\end{bmatrix} \quad (13)$$

Multiplying (13) by the matrix $\begin{bmatrix} 2c-3 & 3c-c^2 \\ 2c+1 & -c-c^2 \end{bmatrix}$ from the left hand side yields the following two equations:

$$\begin{bmatrix} 4c^2(a^2+2c-3) \\ 4c^2(-2ac-c^2+2c+1) \end{bmatrix}=\begin{bmatrix}0\\0\end{bmatrix}. \quad (14)$$

The value $c=0$ is a solution to these equations, which gives $(a^*,c^*)=(0,0)$ as the solution, in which, the corresponding gradient is indefinable. Notice that (14) is also satisfied when $a^2=-2c+3$ and $2ac=-c^2+2c+1$. Merging the two equations, we obtain $c^4+4c^3-10c^2+4c+1=0$. This leads to three solutions $(a^*,c^*)\in\{(1,1),(1-2\sqrt{2},2\sqrt{2}-3),(1+2\sqrt{2},-2\sqrt{2}-3)\}$. In summary we have four solutions: $(a^*,c^*)=(1-2\sqrt{2},2\sqrt{2}-3)$ is the only stationary point inside the feasible set, while $(a^*,c^*)=(0,0)$ is on the boundary. The other two solutions are outside the feasible set. Now it is straightforward to verify that $(1-2\sqrt{2},2\sqrt{2}-3)$ is the global optimal point. We can also demonstrate that the cost function attains a finite value as we take a sequence approaching the stationary point $(a^*,c^*)=(0,0)$ using the stabilizing controller $K_\epsilon=\begin{bmatrix}-\epsilon\\-2\epsilon\end{bmatrix}$. It is clear that $\lim_{\epsilon\to 0^+}J(K_\epsilon)=\frac{1}{2}$ and $\lim_{\epsilon\to 0^+}K_\epsilon\in\partial\mathcal{K}$, illustrating the non-coerciveness of the optimization landscape, thereby completing the proof. ■

IV. Global Convergence of Gradient Flow

In this section, we prove that the policy gradient flow initialized from any point in the feasible set (10) is guaranteed to stay in the feasible set and will converge to the global optimal solution regardless of the lack of coerciveness. For the above example, the gradient flow is defined as

$$\frac{d}{dt}\begin{bmatrix}a(t)\\c(t)\end{bmatrix}=-\nabla J\left(\begin{bmatrix}a(t)\\c(t)\end{bmatrix}\right). \quad (15)$$

The cost value is known to be decreasing along the gradient flow trajectories. However, the gradient flow may not converge to a stationary point due to the possibilities of moving towards infinity or the boundary of the feasible set. Next, we will rule out such possibilities. We will show that from any initial policy in the feasible set, a positive invariant set encompassing the initial policy and the global optimal point (i.e., $\begin{bmatrix}a^*\\c^*\end{bmatrix}=\begin{bmatrix}1-2\sqrt{2}\\2\sqrt{2}-3\end{bmatrix}$) can be formed within the feasible set (10). Consequently, the gradient flow will stay within this set and converge to the unique stationary point in this set, which happens to be the global optimal LQG solution. Let us first establish that the gradient flow never converges to $\begin{bmatrix}0\\0\end{bmatrix}$.

Lemma 2: Let $\begin{bmatrix}a_0\\c_0\end{bmatrix}$ be an initial controller for the gradient flow (15). Define the shaded region in Figure 1 with $m=0.3$ and an $\epsilon>0$ is defined as

$$\epsilon:=\begin{cases}\min\{|a_0|,0.01\} & a_0<0\\ \min\{m(-a_0-c_0),0.01\} & a_0\geq 0.\end{cases}$$

Then the gradient flow does not enter the shaded area.

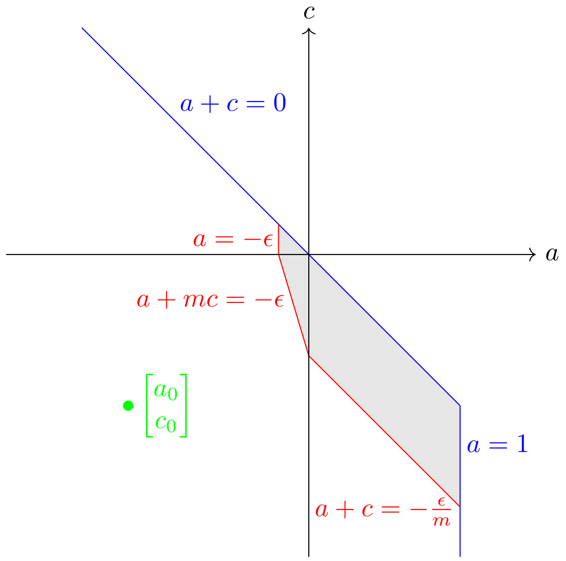
c
$a+c=0$
$a=-\epsilon$
a
$a+mc=-\epsilon$
$\begin{bmatrix}a_0\\c_0\end{bmatrix}$
$a=1$
$a+c=-\frac{\epsilon}{m}$

Fig. 1. Cost Function near point $\begin{bmatrix}0\\0\end{bmatrix}$

Proof: The three line segments that define the shaded area in Fig. 1 are given as: $\mathcal{S}_1:=\{(a,c)\,:\,a=-\epsilon$ and $0\leq c\leq\epsilon\}$, $\mathcal{S}_2:=\{(a,c)\,:\,a+mc=-\epsilon$ and $-\epsilon<a\leq 0\}$, and $\mathcal{S}_3:=\{(a,c)\,:\,a+c=-\frac{\epsilon}{m}$ and $0<a\leq 1\}$. By careful calculations, we can verify that the gradient flow moves outward from the shaded area along these three segments. Due to the page limits, the detailed calculations are given in our full report, which is available at `arXiv`. By choosing ϵ as mentioned in the lemma statement, we can ensure that the initial controller lies outside the shaded region in Fig. 1, and the gradient flow on the boundaries of the shaded regions points outward. This means that even if the gradient flow initialized from the initial controller $\begin{bmatrix}a_0\\c_0\end{bmatrix}$ approaches the boundary of the shaded region, the gradient flow will never enter the shaded area. This gives the desired conclusion. ■

Before proceeding to our construction of the positive invariant set, we need another supporting lemma. Specifically, in the following lemma, we demonstrate that, for any initial controller in the feasible set (10), a subset of the feasible region encompassing both the initial policy and the stationary point can be chosen to ensure that within this subset, the gradient flow remains contained, thereby ruling out the possibility of diverging to infinity.

Lemma 3: Let $\begin{bmatrix}a_0\\c_0\end{bmatrix}$ be an initial controller for the gradient flow (15). There exists a value R_0 (where $R_0\geq\max\{2,\left\|\begin{bmatrix}a_0+c^*\\c_0-c^*\end{bmatrix}\right\|\}$) such that for any circle centered at $\begin{bmatrix}-c^*\\c^*\end{bmatrix}=\begin{bmatrix}3-2\sqrt{2}\\2\sqrt{2}-3\end{bmatrix}$ with radius $R\geq R_0$, the circle encapsulates both the stationary point $\begin{bmatrix}a^*\\c^*\end{bmatrix}$ and the initial policy. Furthermore, when the gradient flow is initialized from this initial policy, it remains within the boundary of the region formed by the intersection of this circle with the boundary region of the feasible set.

Proof: Define the angle associated with a point (a, c) on the circle as $\theta = \text{atan2}\left(\frac{a+c^*}{c-c^*}\right)$. Then each point on the circle can be expressed as:

$$\begin{aligned} a_c(\theta) &= -c^* + R\cos(\theta) \\ c_c(\theta) &= c^* + R\sin(\theta). \end{aligned} \tag{16}$$

As shown in Fig. 2, we split the circle arc into two sections; $\mathcal{S}_1 := \{(a_c(\theta), c_c(\theta)) : \frac{3\pi}{4} < \theta < \frac{3\pi}{2}\}$ and $\mathcal{S}_2 := \{(a_c(\theta), c_c(\theta)) : \frac{3\pi}{2} \leq \theta < \sin^{-1}\left(\frac{1+c^*}{R}\right)\}$. We can establish that the gradient flow points towards the inside of the circle along these two segments. To accomplish this, we need to verify that for every point, the inner product $\begin{bmatrix} \frac{\partial J}{\partial a} \\ \frac{\partial J}{\partial c} \end{bmatrix} \cdot \begin{bmatrix} R\cos(\theta) \\ R\sin(\theta) \end{bmatrix} > 0$.

The verification of the above inner product inequality requires some careful calculations. Such calculations are quite tidious, and definitely non-trivial. Due to the page limit, the detailed calculations for the verification of the above inner product inequality are only given in our full report that is available on `arXiv`. Once the inner produce inequality is verified, then the desired conclusion then directly follows as a consequence. ■

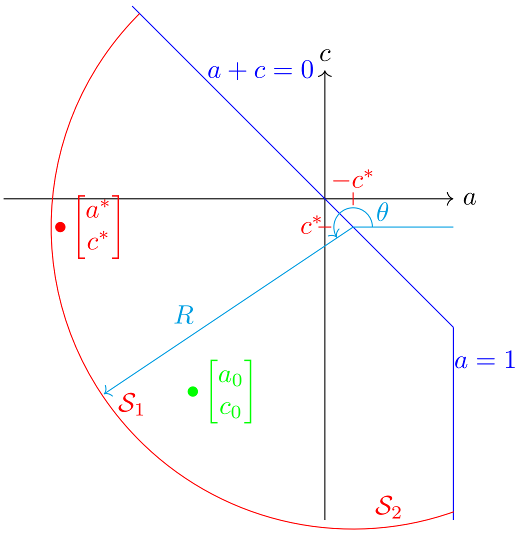
$a + c = 0$
c
$-c^*$
a
θ
c^*
$\begin{bmatrix} a^* \\ c^* \end{bmatrix}$
R
$a = 1$
$\mathcal{S}_1$
$\begin{bmatrix} a_0 \\ c_0 \end{bmatrix}$
$\mathcal{S}_2$

Fig. 2. Cost Function near $\begin{bmatrix} a^* \\ c^* \end{bmatrix}$

Based on Lemma 2 and Lemma 3, we are ready to create a positive invariant set around the initial policy and the stationary point, in which the gradient flow remains confined. This is formalized as follows.

Lemma 4: Let $\left[\begin{smallmatrix} a_0 \\ c_0 \end{smallmatrix}\right]$ be an initial controller for the gradient flow (15). Define a compact set that encapsulates both the initial policy point $\left[\begin{smallmatrix} a_0 \\ c_0 \end{smallmatrix}\right]$ and the stationary point $\left[\begin{smallmatrix} a^* \\ c^* \end{smallmatrix}\right] = \begin{bmatrix} 1-2\sqrt{2} \\ 2\sqrt{2}-3 \end{bmatrix}$. This set is structured into six segments, labeled as $\mathcal{S}_1$, $\mathcal{S}_2$, $\mathcal{S}_3$, $\mathcal{S}_4$, $\mathcal{S}_5$, and $\mathcal{S}_6$, as depicted in Figure IV. The delineation of these segments is as follows:

$$\begin{aligned}
\mathcal{S}_1 &= \{(a, c) : a + c = -\epsilon_1 \text{ and } -c^* - \frac{R}{\sqrt{2}} \leq a \leq -\epsilon_2\} \\
\mathcal{S}_2 &= \{(a, c) : a = -\epsilon_2 \text{ and } 0 \leq c \leq \epsilon_2 - \epsilon_1\} \\
\mathcal{S}_3 &= \{(a, c) : a + mc = -\epsilon_2 \text{ and } -\epsilon_2 \leq a \leq 0\} \\
\mathcal{S}_4 &= \{(a, c) : a + c = -\frac{\epsilon_2}{m} \text{ and } 0 < a \leq 1 - \epsilon_3\} \\
\mathcal{S}_5 &= \{(a, c) : a = 1 - \epsilon_3 \text{ and} \\
&\quad c^* - R\cos(\theta_1) \leq c \leq \epsilon_3 - \frac{\epsilon_2}{m} - 1\} \\
\mathcal{S}_6 &= \{(a_c(\theta), c_c(\theta)) : \frac{3\pi}{4} < \theta < \frac{3\pi}{2} + \theta_1\}
\end{aligned}$$

where $m = 0.3$, ϵ_1, ϵ_2 and ϵ_3 are positive real numbers which are defined as

$$\begin{aligned}
\epsilon_2 &:= \begin{cases} \min\{|a_0|, 0.01\} & a_0 < 0 \\ \min\{m(-a_0 - c_0), 0.01\} & a_0 \geq 0 \end{cases} \\
\epsilon_1 &\in (0, \epsilon_2] \\
\epsilon_3 &\in \begin{cases} (0, 1 - a_0] & a_0 \geq 0 \\ (0, 1) & a_0 < 0 \end{cases}
\end{aligned}$$

Also, R is defined as lemma 3 and $\theta_1 = \sin^{-1}\left(\frac{1+c^*-\epsilon_3}{R}\right)$. The points $(a_c(\theta), c_c(\theta))$ on the circle $\mathcal{S}_6$ are defined as in (16). The gradient flow initialized from $\left[\begin{smallmatrix} a_0 \\ c_0 \end{smallmatrix}\right]$ remains within this compact set, which establishes this compact set a positive invariant set.

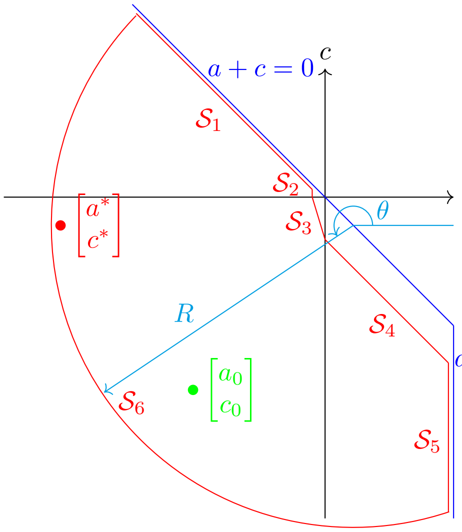

Fig. 3. Cost Function far from $\begin{bmatrix} a_0 \\ c_0 \end{bmatrix}$

Proof: We first establish the motion of the gradient flow towards the interior of the compact set along all six segments. Our previous analysis, as shown in Lemma 2, confirms the inward movement of the gradient flow on $\mathcal{S}_2$, $\mathcal{S}_3$ and $\mathcal{S}_4$. Furthermore, Lemma 3 demonstrates the inward direction of the gradient flow along $\mathcal{S}_6$. To complete the analysis, we must establish that the gradient flow, for every point along $\mathcal{S}_1$ and $\mathcal{S}_5$, also moves towards the inside of the compact set.

Analysis along $\mathcal{S}_1$: Here we need to show that for every point on $\mathcal{S}_1$, $-\begin{bmatrix}\frac{\partial J}{\partial a}\\ \frac{\partial J}{\partial c}\end{bmatrix}\cdot\begin{bmatrix}-1\\-1\end{bmatrix} > 0$. From (12), we have

$$\begin{aligned}-\begin{bmatrix}\frac{\partial J}{\partial a}\\ \frac{\partial J}{\partial c}\end{bmatrix}\cdot\begin{bmatrix}-1\\-1\end{bmatrix} &= \frac{\partial J}{\partial a}+\frac{\partial J}{\partial c}\\ &= \frac{(-2c-1)a^3+(-3c^2+9c+1)a^2+(c^2-6c)a}{2(a+c)^2(a-1)^2}\\ &+\frac{c^4-c^3+c^2}{2(a+c)^2(a-1)^2}\end{aligned} \tag{17}$$

The denominator of (17) is positive along this line segment; therefore, we only need to demonstrate that the numerator is also positive. Substituting $c=-a-\epsilon_1$ into the numerator yields the following to prove

$$\begin{aligned}&-8a^3+(3\epsilon_1^2-4\epsilon_1+8)a^2+(4\epsilon_1^3+4\epsilon_1^2+8\epsilon_1)a\\ &+\epsilon_1^4+\epsilon_1^3+\epsilon_1^2>0.\end{aligned} \tag{18}$$

To establish the validity of (18), we aim to demonstrate its applicability for all values of a within the range $a\le-\epsilon_2\le-\epsilon_1$. Given from Lemma 2 that ϵ_1 is constrained to be less than or equal to 0.01, we can divide this proof into two distinct cases:

Case 1: $-1\le a\le-\epsilon_1$:
In this case, we establish a lower bound for (18) as follows:

$$\begin{aligned}&-8a^3+(e\epsilon_1^2-4\epsilon_1+8)a^2+(4\epsilon_1^3+4\epsilon_1^2+8\epsilon_1)a\\ &+\epsilon_1^4+\epsilon_1^3+\epsilon_1^2>^{(i)}(4\epsilon_1+8)a^2+(\epsilon_1^3+4\epsilon_1^2+8\epsilon_1)a\\ &+\epsilon_1^4+\epsilon_1^3+\epsilon_1^2>^{(ii)}\epsilon_1^4+(1+a)\epsilon_1^3+\epsilon_1^2>^{(iii)}0,\end{aligned} \tag{19}$$

where (i) holds because $-8a^3\ge8\epsilon_1a^2$ and $3\epsilon_1^2a^2\ge-3\epsilon_1^3a$. Inequality (ii) holds due to $4\epsilon_1a^2 \ge -4\epsilon_1^2a$ and $8a^2\ge-8\epsilon_1a$. Finally, (iii) follows from $-1\le a$.

Case 2: $a\le-1$:
In this case, we establish a lower bound for (18) as follows:

$$\begin{aligned}&-8a^3+(e\epsilon_1^2-4\epsilon_1+8)a^2+(4\epsilon_1^3+4\epsilon_1^2+8\epsilon_1)a\\ &+\epsilon_1^4+\epsilon_1^3+\epsilon_1^2>^{(i)}(4\epsilon_1+8)a^2+(\epsilon_1^3+4\epsilon_1^2+8\epsilon_1)a\\ &+\epsilon_1^4+\epsilon_1^3+\epsilon_1^2>^{(ii)}8a(a+\frac{13}{8}\epsilon_1)>^{(iii)}0\end{aligned} \tag{20}$$

Inequality (i) holds because $-8a^3\ge8\epsilon_1a^2$ and $3\epsilon_1^2a^2\ge-3\epsilon_1^3a$. Inequality (ii) is true as $\epsilon_1\ge\epsilon_1^2\ge\epsilon_1^3>0$. Inequality (iii) holds due to our assumption that $\epsilon_1\le0.01$ and $a\le-1$. Thus, we show that for every points along along $\mathcal{S}_1$, $-\begin{bmatrix}\frac{\partial J}{\partial a}\\ \frac{\partial J}{\partial c}\end{bmatrix}\cdot\begin{bmatrix}-1\\-1\end{bmatrix}>0$.
Analysis along $\mathcal{S}_5$: Here we need to show that for every point on $\mathcal{S}_5$, $-\begin{bmatrix}\frac{\partial J}{\partial a}\\ \frac{\partial J}{\partial c}\end{bmatrix}\cdot\begin{bmatrix}-1\\0\end{bmatrix}>0$. From (12), we have

$$\begin{aligned}-\begin{bmatrix}\frac{\partial J}{\partial a}\\ \frac{\partial J}{\partial c}\end{bmatrix}\cdot\begin{bmatrix}-1\\0\end{bmatrix} &= \frac{\partial J}{\partial a}\\ &= \frac{1+c^2}{2(a-1)^2}+\frac{-a^2-3c^2}{2(a-1)^2(a+c)}\\ &+\frac{-a^2-3c^2}{2(a-1)(a+c)^2}+\frac{a}{(a+c)(a-1)}\\ &>^{(i)}0\end{aligned} \tag{21}$$

where inequality (i) holds since along $\mathcal{S}_5$, $a=1-\epsilon_3>0$ and thus all the terms in (21) are positive. Consequently, we guarantee that $-\begin{bmatrix}\frac{\partial J}{\partial a}\\ \frac{\partial J}{\partial c}\end{bmatrix}\cdot\begin{bmatrix}-1\\0\end{bmatrix}>0$ for every point along $\mathcal{S}_5$. By choosing ϵ_1 and ϵ_2, ϵ_3 and R as mentioned before, we ensure that the initial controller lies inside the compact set and the gradient flow on the boundaries of the compact set points inward. This implies that even if the initial controller $\begin{bmatrix}a_0\\c_0\end{bmatrix}$ approaches the boundaries, the gradient flow points inward the compact set, demonstrating its positive invariance nature. The invariance can be affirmed using, for instance, Nagumo's theorem [38], [39] or direct methods, thereby completing the proof. ■

Now, we are ready to demonstrate the convergence of gradient flow to the global stationary point.

Theorem 1: Starting from any initial controller $\begin{bmatrix}a_0\\c_0\end{bmatrix}$ within feasible set (10), the gradient flow (15) will stay in the feasible set and converge to the stationary point $\begin{bmatrix}a^*\\c^*\end{bmatrix}=\begin{bmatrix}1-2\sqrt{2}\\2\sqrt{2}-3\end{bmatrix}$, which is the unique global optimal point.

Proof: Given any initial controller $\begin{bmatrix}a_0\\c_0\end{bmatrix}$, we can apply Lemma 4 to construct a compact set encompassing both the initial policy and the stationary point. Furthermore, it is established that this compact set $\mathcal{S}$ is positive invariant with respect to (15). Consider the Lyapunov function $V=J$. From (15), we know that $\dot{V}\le0$. Since $\begin{bmatrix}a^*\\c^*\end{bmatrix}$ is the only point where $\dot{V}=0$, we can use LaSalle's theorem [40] to show that the gradient flow approaches $\begin{bmatrix}a^*\\c^*\end{bmatrix}$ as $t\to\infty$. This completes the proof. ■

V. Conclusion

In this paper, we delved into LQG policy optimization by examining a specific case of a single-input single-output (SISO) system. To prove the global convergence of direct policy search on this problem, we need to overcome two main difficulties. Firstly, to remove suboptimal stationary points, we adopt a controllable policy parameterization approach. Secondly, we need to address the non-coerciveness of the LQG cost function. Our resolution to this challenge was to create a positive invariant set within the feasible set, which includes both the initial policy and the global stationary point. Specifically, for this SISO example, our approach in constructing the positive invariant set allows us to prove that the policy gradient flow is guaranteed to stay in the feasible set and converge to the global optimal solution regardless of the non-coerciveness issue. It is our hope that the understandings derived from this case study could illuminate broader LQG policy optimization problems.

VI. Acknowledgement

The work of Darioush Kevian and Geir Dullerud was supported by NSF under Grant ECCS 1932735. The work of Peter Seiler was supported by the U.S. Office of Naval Research (ONR) under Grant N00014-18-1-2209. The work of Bin Hu is generously supported by the NSF award CAREER-2048168.

References

[1] 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.

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

[3] B. Hu, K. Zhang, N. Li, M. Mesbahi, M. Fazel, and T. Başar, "Toward a theoretical foundation of policy optimization for learning control policies," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 6, pp. 123–158, 2023.

[4] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator," in *International conference on machine learning*. PMLR, 2018, pp. 1467–1476.

[5] 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.

[6] D. Malik, A. Pananjady, K. Bhatia, K. Khamaru, P. Bartlett, and M. Wainwright, "Derivative-free methods for policy optimization: Guarantees for linear quadratic systems," in *The 22nd international conference on artificial intelligence and statistics*. PMLR, 2019, pp. 2916–2925.

[7] H. Mohammadi, A. Zare, M. Soltanolkotabi, and M. R. Jovanović, "Convergence and sample complexity of gradient methods for the model-free linear–quadratic regulator problem," *IEEE Transactions on Automatic Control*, vol. 67, no. 5, pp. 2435–2450, 2021.

[8] Z. Jin, J. M. Schmitt, and Z. Wen, "On the analysis of model-free methods for the linear quadratic regulator," *arXiv preprint arXiv:2007.03861*, 2020.

[9] Z. Yang, Y. Chen, M. Hong, and Z. Wang, "Provably global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost," *Advances in neural information processing systems*, vol. 32, 2019.

[10] B. Hambly, R. Xu, and H. Yang, "Policy gradient methods for the noisy linear quadratic regulator over a finite horizon," *SIAM Journal on Control and Optimization*, vol. 59, no. 5, pp. 3359–3391, 2021.

[11] K. Zhang, B. Hu, and T. Basar, "Policy optimization for $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$ robustness guarantee: Implicit regularization and global convergence," *SIAM Journal on Control and Optimization*, vol. 59, no. 6, pp. 4081–4109, 2021.

[12] K. Zhang, X. Zhang, B. Hu, and T. Basar, "Derivative-free policy optimization for linear risk-sensitive and robust control design: Implicit regularization and sample complexity," *Advances in neural information processing systems*, vol. 34, pp. 2949–2964, 2021.

[13] 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, 2020.

[14] F. Zhao and K. You, "Primal-dual learning for the model-free risk-constrained linear quadratic regulator," in *Learning for Dynamics and Control*, 2021, pp. 702–714.

[15] Y. Zhang, Z. Yang, and Z. Wang, "Provably efficient actor-critic for risk-sensitive and robust adversarial RL: A linear-quadratic case," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 2764–2772.

[16] K. Zhang, Z. Yang, and T. Basar, "Policy optimization provably converges to nash equilibria in zero-sum linear quadratic games," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[17] J. Bu, L. J. Ratliff, and M. Mesbahi, "Global convergence of policy gradient for sequential zero-sum linear quadratic dynamic games," *arXiv preprint arXiv:1911.04672*, 2019.

[18] 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, pp. 22 056–22 068, 2020.

[19] 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)*. IEEE, 2020, pp. 2882–2887.

[20] 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.

[21] J. P. Jansch-Porto, B. Hu, and G. E. Dullerud, "Policy optimization for markovian jump linear quadratic control: Gradient method and global convergence," *IEEE Transactions on Automatic Control*, vol. 68, no. 4, pp. 2475–2482, 2022.

[22] J. Perdomo, J. Umenberger, and M. Simchowitz, "Stabilizing dynamical systems via policy gradient methods," *Advances in neural information processing systems*, vol. 34, pp. 29 274–29 286, 2021.

[23] I. K. Ozaslan, H. Mohammadi, and M. R. Jovanović, "Computing stabilizing feedback gains via a model-free policy gradient method," *IEEE Control Systems Letters*, vol. 7, pp. 407–412, 2022.

[24] F. Zhao, X. Fu, and K. You, "On the sample complexity of stabilizing linear systems via policy gradient methods," *arXiv preprint arXiv:2205.14335*, 2022.

[25] X. Guo and B. Hu, "Global convergence of direct policy search for state-feedback $\mathcal{H}_\infty$ robust control: A revisit of nonsmooth synthesis with goldstein subdifferential," in *36th Conference on Neural Information Processing Systems*, vol. 28, 2022.

[26] H. Feng and J. Lavaei, "On the exponential number of connected components for the feasible set of optimal decentralized control problems," in *2019 American Control Conference (ACC)*. IEEE, 2019, pp. 1430–1437.

[27] 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.

[28] Y. Tang and Y. Zheng, "On the global optimality of direct policy search for nonsmooth $\mathcal{H}_\infty$ output-feedback control," in *2023 62nd IEEE Conference on Decision and Control (CDC)*, 2023, pp. 6148–6153.

[29] J. Umenberger, M. Simchowitz, J. Perdomo, K. Zhang, and R. Tedrake, "Globally convergent policy search for output estimation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 22 778–22 790, 2022.

[30] J. Duan, W. Cao, Y. Zheng, and L. Zhao, "On the optimization landscape of dynamic output feedback linear quadratic control," *IEEE Transactions on Automatic Control*, 2023.

[31] X. Zhang, B. Hu, and T. Başar, "Learning the kalman filter with fine-grained sample complexity," in *2023 American Control Conference (ACC)*, 2023, pp. 4549–4554.

[32] X. Guo, D. Keivan, G. Dullerud, P. Seiler, and B. Hu, "Complexity of derivative-free policy optimization for structured $\mathcal{H}_\infty$ control," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[33] Y. Zheng, Y. Tang, and N. Li, "Analysis of the optimization landscape of linear quadratic gaussian LQG control," *arXiv preprint arXiv:2102.04393*, 2021.

[34] B. Hu and Y. Zheng, "Connectivity of the feasible and sublevel sets of dynamic output feedback control with robustness constraints," *IEEE Control Systems Letters*, vol. 7, pp. 442–447, 2022.

[35] Y. Zheng, Y. Sun, M. Fazel, and N. Li, "Escaping high-order saddles in policy optimization for linear quadratic gaussian (LQG) control," in *2022 IEEE 61st Conference on Decision and Control (CDC)*. IEEE, 2022, pp. 5329–5334.

[36] Y. Zheng, C.-f. Pai, and Y. Tang, "Benign nonconvex landscapes in optimal and robust control, part i: Global optimality," *arXiv preprint arXiv:2312.15332*, 2023.

[37] K. Zhou, J. C. Doyle, and K. Glover, *Robust and Optimal Control*. Prentice Hall New Jersey, 1996, vol. 40.

[38] M. Nagumo, "Über die lage der integralkurven gewöhnlicher differentialgleichungen," *Proceedings of the Physico-Mathematical Society of Japan. 3rd Series*, vol. 24, pp. 551–559, 1942.

[39] F. Blanchini, S. Miani *et al.*, *Set-theoretic methods in control*. Springer, 2008, vol. 78.

[40] K. K. Hassan *et al.*, "Nonlinear systems," *Departement of Electrical and computer Engineering, Michigan State University*, 2002.