Global Convergence of Policy Gradient Methods for Output Feedback Linear Quadratic Control

Feiran Zhao, Xingyun Fu, Keyou You

Department of Automation and BNRist, Tsinghua University, Beijing 100084, China (e-mail: zhaofr18@mails.tsinghua.edu.cn; fjy20@mails.tsinghua.edu.cn; youky@tsinghua.edu.cn.)

Abstract: While the optimization landscape of policy gradient methods has been recently investigated for partially observable linear systems in terms of both dynamical controllers and static output feedback, they can only provide convergence guarantees to stationary points. In this paper, we propose a new parameterization of the policy, which uses a past input-output trajectory of finite length as the feedback. We show that the solution set to the parameterized optimization problem is a matrix space, which is invariant to similarity transformation. By proving a gradient dominance property, we show the global convergence of policy gradient methods. Moreover, we observe that the gradient is orthogonal to the solution set, revealing an explicit relation between the resulting solution and the initial policy. Finally, we perform simulations to validate our theoretical results.

Keywords: Data-driven control, Linear quadratic regulator, Output feedback control, Reinforcement learning, Optimal control.

1. INTRODUCTION

Recent years have witnessed tremendous successes of reinforcement learning (RL) in applications such as sequential decision-making problems (Mnih et al., 2015; Silver et al., 2016) and continuous control (Tobin et al., 2017; Levine et al., 2016; Andrychowicz et al., 2020; Recht, 2019). As an essential approach of RL, the policy gradient (PG) method directly searches over a policy space to optimize a performance index of interests by using sampled trajectories, and thus does not requires any identification process. Such an end-to-end approach is conceptually simple and easy to implement in practice.

Compared to the above empirical successes, the theoretical understanding of the PG method has largely lagged as it often involves challenging non-convex optimization problems. To fill this gap, there has been a resurgent interest in studying the theoretical properties of PG methods for classical control problems (Fazel et al., 2018; Gravell et al., 2020; Zhao et al., 2022, 2021; Malik et al., 2019; Zhang et al., 2021; Li et al., 2021; Zheng et al., 2021, 2022; Duan et al., 2022a,b; Fatkhullin and Polyak, 2021). The seminal work of Fazel et al. (2018) shows that the LQR cost has a gradient dominance property with respect to the gain matrix, which leads to the global convergence of PG methods despite the non-convexity. There are also other PG-based works considering, e.g., system stabilization (Zhao et al., 2022), robustness (Zhang et al., 2021) and distributed control (Li et al., 2021), just to name a few. For a comprehensive overview, we refer the readers to the survey (Hu et al., 2022).

In this paper, we consider partially observable linear systems, where the state is unmeasurable and only past input-output trajectories can be used for feedback control. Recently, there have been several works studying the PG methods for the output feedback control (Zheng et al., 2022, 2021; Duan et al., 2022a,b; Fatkhullin and Polyak, 2021). Depending on how the input-output information is used, they can be broadly categorized into static output feedback (SOF) (Fatkhullin and Polyak, 2021; Duan et al., 2022b) and dynamic output feedback (Zheng et al., 2021, 2022; Duan et al., 2022a). The former class only uses the current output as feedback, while the latter uses all past input-output trajectories by invoking a linear filter. In both classes, the optimization landscape of PG methods can be substantially different from that in state-feedback control. Particularly, the gradient dominance property does not hold, which is the key to the convergence of the LQR (Fazel et al., 2018). Moreover, the set of stabilizing controllers is usually disconnected, and stationary points can be local minima or saddle points (Fatkhullin and Polyak, 2021; Duan et al., 2022b). Even though a perturbed PG method is proposed to escape the strict saddle, its convergence rate has not been well characterized yet (Zheng et al., 2022, 2021). Last but not least, the cost function may vary with similarity transformations (Duan et al., 2022a), which further increases the difficulty in the convergence analysis. Therefore, all the above works for output feedback can only provide convergence guarantees to stationary points.

This paper provides the first global convergence result of PG methods for output feedback control. We propose a new policy parameterization in the form of input-output feedback (IOF), which uses a past input-output trajectory...
of fixed length instead of the current output. We show that the solution set to the parameterized optimization problem is a matrix space, which is invariant to the similarity transformation. Interestingly, though an optimal policy is not unique, our problem still enjoys the gradient dominance property, which shows the merits of our parameterization. Based on the above theoretical findings, we prove the global convergence of the PG method. Moreover, we reveal an explicit relation between the converged solution and the initial policy by observing that the gradient is orthogonal to the solution set. Finally, we propose a zero-order algorithm with warm-up cost evaluation for sample-based implementation.

The remainder of this paper is organized as follows. Section 2 formulates the output feedback linear quadratic control problem as a parameterized optimization problem. Section 3 derives its optimal solution which is shown to be invariant to similarity transformation. Section 4 shows the convergence of the PG method. Section 5 discusses its implementation in the sample-based setting. Section 6 performs simulations to validate our results. Finally, the conclusion is made in Section 7.

2. PROBLEM FORMULATION

We consider the following standard discrete-time linear time-invariant system

\[
x_{t+1} = Ax_t + Bu_t, \quad y_t = Cx_t,
\]

where \(x_t \in \mathbb{R}^n\) is the state, \(u_t \in \mathbb{R}^m\) is the control input, and \(y_t \in \mathbb{R}^q\) is the measurable output. The matrices \(A \in \mathbb{R}^{n \times n}, B \in \mathbb{R}^{n \times m}\) and \(C \in \mathbb{R}^{d \times n}\) are model parameters.

We aim to find a policy sequence \(\{\pi_t\}\) using only past input-output data to minimize an infinite-horizon quadratic cost, i.e.,

\[
\begin{aligned}
\min_{\{\pi_t\}} & \quad \mathbb{E}_x \left[ \sum_{t=0}^{\infty} y_t^\top Q y_t + u_t^\top R u_t \right] \\
\text{subject to} & \quad (1), x_0 \sim \mathcal{D} \quad \text{and} \\
& \quad u_t = \pi_t(u_{t-1}, y_{t-1}, \ldots, y_{t-L})
\end{aligned}
\]

with \(Q \geq 0, R > 0\). We require the distribution \(\mathcal{D}\) of the initial state to satisfy the following assumption.

Assumption 1. \(\mathcal{D}\) has zero mean with some positive definite covariance matrix \(\Sigma_0 = \mathbb{E}[x_0 x_0^\top] > 0\).

In (2), we allow the controller to use all past input-output information. Since the controller does not depend on the current output \(y_t\), it can be applied to strictly causal systems. We make the following assumption standard in the control theory (Bertsekas, 2012) such that (1) is a minimal realization.

Assumption 2. The pair \((A, B)\) is controllable, \((CQ^{1/2}, A)\) and \((C, A)\) are observable.

In contrast to Duan et al. (2022a,b), we do not require the observation matrix \(C\) to have full row rank (Duan et al., 2022a) or full column rank (Duan et al., 2022b).

It is well-known that when the state is measurable, an optimal policy is a state linear feedback

\[
u_t = -(R + B^\top P B)^{-1} B^\top P^* A x_t,
\]

where \(P^*\) is a unique positive semi-definite solution to the Algebraic Riccati Equation (ARE)

\[
P^* = A^\top P^* A + Q_c - A^\top P^* B (R + B^\top P B)^{-1} B^\top P^* A + Q_c C^\top QC.
\]

Note that the optimal policy (3) is not related to \(\Sigma_0\).

Since \(x\) is unmeasurable in (2), this paper considers the following policy parameterization

\[
u_t = -K z_{t,p},
\]

where \(z_{t,p} = [y_{t-1, p}, \ldots, y_{t,p}^\top]^\top\), \(u_{t,p}\) and \(y_{t,p}\) are past input-output data defined by \(u_{t,p} = [u_{t-1, p}, \ldots, u_{t,p}^\top]^\top\) and \(y_{t,p} = [y_{t-1, p}, \ldots, y_{t,p}^\top]^\top\), and \(p \in \mathbb{N}\) is a system-dependent constant to be defined later. The gain matrix \(K \in \mathbb{R}^{m \times q}\) with \(q = p(m + d)\) is the policy parameter to be optimized by gradient methods. That is, we solve the following optimization problem by viewing \(K\) as the variable

\[
\min_K J(K) := \mathbb{E}_x \left[ \sum_{t=0}^{\infty} y_t^\top Q y_t + u_t^\top R u_t \right]
\]

subject to (1), (4) and \(x_0 \sim \mathcal{D}\).

The intuition behind the parameterization in (4) is that the state can be recovered under Assumption 2 by a finite-length past input-output trajectory. In this paper, we solve the optimal policy in the form of (4) and study its optimization landscape.

3. OPTIMAL OUTPUT FEEDBACK CONTROLLER

In this section, we show that the solution set to (5) is a matrix space. Moreover, it is invariant to the similarity transformation.

Let \(o\) and \(c\) be the observability index and controllability index, respectively, and \(p = \max(o, c) \leq n\). Then, the following matrices

\[
\mathcal{O}_p = \begin{bmatrix} CA^{p-1} \\ \vdots \\ CA \\ C \end{bmatrix}, \quad \text{and} \quad \mathcal{C}_p = [B \ AB \ \cdots \ A^{p-1}B]
\]

have full column and row rank, respectively. At time step \(t\), the state can be represented using system dynamics and history trajectories as

\[
x_{t} = A^p x_{t-p} + \mathcal{C}_p u_{t-p},
\]

\[
y_t = \mathcal{O}_p x_{t-p} + \mathcal{T}_p u_{t-p}
\]

with a Toeplitz matrix

\[
\mathcal{T}_p = \begin{bmatrix} 0 & CB & CAB & \cdots & CA^{p-2}B \\ 0 & 0 & CB & \cdots & CA^{p-3}B \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & CB \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}
\]

In some cases, \(p\) is unknown, and we only have knowledge of the system order \(o\), input dimension \(m\), and output dimension \(d\). Then, one can substitute \(p\) with \(n\) in (6), and the results in this paper still hold. For simplicity, we omit the subscript \(p\) where it can be understood from the context.

Since \(\mathcal{O}\) has full column rank, it has a unique left pseudo inverse \(\mathcal{O}^\dagger = (\mathcal{O}^\top \mathcal{O})^{-1} \mathcal{O}^\top\). Then, it follows immediately
from (7) that \( x_t \) can be uniquely determined by eliminating \( x_{t-p} \) as
\[
x_t = (\mathcal{C} - A^pO^T)u_{t,p} + A^pO^Ty_{t,p} := Sx_{t,p},
\]
with \( S = [\mathcal{C} - A^pO^T, A^pO^T] \). Clearly, \( S \) has full row rank by noting
\[
S \begin{bmatrix} I & 0 \\ T & I \end{bmatrix} = [\mathcal{C} - A^pO^T, A^pO^T] \begin{bmatrix} I & 0 \\ T & I \end{bmatrix} = [\mathcal{C}, A^pO^T],
\]
and has a unique right pseudo inverse \( S^\dagger = S^\top (SS^\top)^{-1} \). Denote the feasible set of \( K \) as
\[
S = \{ K \in \mathbb{R}^{m \times q} | \rho(A - BK S^\dagger) < 1 \}.
\]
We have the following closed-form expression for \( J(K) \).

**Lemma 3.** For any \( K \in S \), the cost function \( J(K) \) can be written as
\[
J(K) = \text{Tr}(P_K \Sigma_0),
\]
where \( P_K \geq 0 \) is the solution to the Lyapunov equation
\[
P_K = Q + (S^\top)^\top F \top RK S^\dagger
+ (A - BK S^\dagger)^\top P_K (A - BK S^\dagger).
\]

**Proof.** Let \( V_K(x) = x^\top Px \) be the value function of problem (5) following the stabilizing policy \( K \). By the well-known Bellman equation (Bertsekas, 2012), it follows that
\[
V_K(x_t) = y_t^\top Qy_t + (-Kz_{t,p})^\top R(-Kz_{t,p}) + V_K(x_{t+1}).
\]
Then, substituting \( x_t \) with (8) yields that
\[
z_{t,p}^\top S^\top PSz_{t,p} = z_{t,p}^\top S^\top Qz_{t,p} + z_{t,p}^\top K^\top RKz_{t,p}
+ z_{t,p}^\top (AS - BK)^\top P(AS - BK)z_{t,p}.
\]
Noting that it holds for all \( z_{t,p} \), it holds that
\[
S^\top PS = S^\top Q_s S + K^\top RK + (AS - BK)^\top P(AS - BK).
\]
Pre- and post-multiplying \( (S^\top)^\top \) and \( S^\top \) in both sides of the above equation yields (10). \( \Box \)

Clearly, an optimal policy of form (4) can be determined by substituting (8) into (3) as
\[
u_t = -K^* z_{t,p},
\]
with
\[
K^* = (R + B^\top P^* B)^{-1} B^\top P^* AS
\]
which also satisfies the following Lyapunov equation
\[
P^* = Q + (S^\top)^\top (K^*)^\top RK S^\dagger
+ (A - BK S^\dagger)^\top P^* (A - BK S^\dagger).
\]
However, an optimal solution is not unique as \( S^\dagger \) does not have full row rank. Define the matrix space
\[
\mathcal{F} = \{ \Delta \in \mathbb{R}^{m \times q} | \Delta \cdot S^\dagger = 0 \}.
\]
We show in the following theorem that the solution set to (5) is a matrix space parallel to \( \mathcal{F} \).

**Theorem 4.** Define the set \( \mathcal{K} = \{ K \in \mathbb{R}^{m \times q} | K = K^* + \Delta, \Delta \in \mathcal{F} \} \). Then, \( K \) is an optimal policy to (5) if and only if \( K \in \mathcal{K} \).

**Proof.** To prove the “if” statement, let \( K \in \mathcal{K} \). By the definition of \( \mathcal{K} \), it holds that \( K S^\dagger = (K^* + \Delta) S^\dagger = K^* S^\dagger + \Delta \in \mathcal{F} \). Combining with (10), all \( K \in \mathcal{K} \) have the same optimal cost as \( K^* \), which implies that \( K \in \mathcal{K} \) is optimal.

To prove the “only if” statement, suppose that \( K \) is optimal, i.e., \( K \) satisfies
\[
P^* = Q + (S^\top)^\top K^\top RK S^\dagger + (A - BK S^\dagger)^\top P^* (A - BK S^\dagger).
\]
Let \( K S^\dagger = K^* S^\dagger + \Delta \). Taking \( K S^\dagger \) into the above equation leads to that
\[
(K^* S^\dagger)^\top RE + E^\top RK^* S^\dagger + E^\top (R + B^\top P^* B)E
= (A - BK^* S^\dagger)^\top P^* BE + (BE)^\top P^* (A - BK^* S^\dagger).
\]
Then, inserting (12) into it yields that \( E^\top (R + B^\top P^* B)E = 0 \). Thus, we can only have \( E = 0 \), i.e., \( K \in \mathcal{K} \). The proof is now completed. \( \Box \)

At the first sight, Theorem 4 appears to be a negative result, as the global convergence of PG methods has been shown in the existing literature only when an optimal policy is unique. In fact, the convergence in our problem can be proved by utilizing the special structure of the solution set, as to be shown in Section 4.

Finally, we show that the optimal gain in (11) is invariant to similarity transformations.

**Lemma 5.** For a nonsingular matrix \( T \), define the new system with the similarity transformation \( \tilde{x}_t = Tx_t \)
\[
\tilde{x}_{t+1} = TA^{-1}\tilde{x}_t + TBu_t,
\]
\[
y_t = CT^{-1}\tilde{x}_t.
\]
Then, \( K \) is an optimal policy of (14) if and only if \( K \in \mathcal{K} \).

**Proof.** The ARE of (14) can be written as
\[
0 = (TA^{-1})^\top \tilde{P}^* T A^{-1} - \tilde{P}^* + (CT^{-1})^\top QC T^{-1}
- (TA^{-1})^\top \tilde{P}^* T B (R + (TB)^\top \tilde{P}^* T B)^{-1} (TB)^\top \tilde{P}^* T A^{-1}.
\]
Pre- and post-multiplying \( T^\top \) and \( T \), it follows that \( \tilde{P}^* = (T^{-1})^\top P^* T^\top \). Similarly, we can show that \( \tilde{S} = TS \).

Hence, an optimal gain matrix of the new system is
\[
\tilde{K}^* = (R + (TB)^\top \tilde{P}^* T B)^{-1} (TB)^\top \tilde{P}^* T A^{-1} \tilde{S}
= (R + B^\top P^* B)^{-1} B^\top P^* AS
= K^*.
\]
Noting that \( \mathcal{N}(\tilde{S})^\top = \mathcal{N}(S^\top)^\top \), the proof is completed. \( \Box \)

**Remark 1.** The optimal dynamical controller in Duan et al. (2022a); Zheng et al. (2021) has the following form
\[
\dot{\xi} = (A - BK)\xi + L(y - C\xi)
\]
\[
u = -K\xi,
\]
where \( K \) is the LQR gain and \( L \) is the Kalman gain. Clearly, it is not unique and each similarity transformation leads to a different optimal policy, which makes it much more challenging to provide global convergence guarantees.

Lemma 5 implies that we can focus on the minimal realization in (1) to study the optimization landscape and the convergence of the PG method.

4. OPTIMIZATION LANDSCAPE

In this section, we show the global convergence of PG methods by proving a gradient dominance property. More-
Theorem 9. For an appropriate stepsize \( \eta \) such that the gradient update (16) converges to the Lyapunov function \( \Sigma_K = \Sigma_0 + (A - BS)S(A - BS)^\top \). Then, we have the following gradient expression.

**Lemma 6.** For \( K \in \mathcal{S} \), the gradient of \( J(K) \) is \( \nabla J(K) = 2E_K\Sigma_K(S^\top)^\top \), where \( E_K = (R + B^\top P_KB)KS^\top - B^\top P_KA \).

**Proof.** By Fazel et al. (2018, Theorem 1), the gradient with respect to \( X \) can be written as \( \nabla_X J = 2E_K\Sigma_K \).

Then, it follows from the chain rule that \( \nabla J(K) = \nabla_X J \cdot (S^\top)^\top = 2E_K\Sigma_K(S^\top)^\top \).

Consider the following gradient method to update \( K \):

\[
K^{i+1} = K^i - \eta \nabla J(K^i), \quad i \in \{0,1,\ldots\}
\]

As in the standard LQR (Fazel et al., 2018), we observe that the gradient dimensionality property (aka Polyak-Lojasiewicz condition (Polyak, 1963)), which guarantees that all stationary points are optimal.

**Lemma 7.** For any \( K \in \mathcal{S} \), it holds that

\[
J(K) - J(K^*) \leq \frac{\left\| \Sigma_K \right\| \left\| S \right\|^2}{4\alpha(R)\sigma^2(\Sigma_K)} \text{tr}\{\nabla J(K)\top \nabla J(K)\}.
\]

**Proof.** By Fazel et al. (2018, Corollary 5), we can show that the cost satisfies

\[
J(K) - J(K^*) \leq \frac{\left\| \Sigma_K \right\|}{4\alpha(R)\sigma^2(\Sigma_K)} \text{tr}\{\nabla X J^\top \nabla X J\}
\]

\[
= \frac{\left\| \Sigma_K \right\|}{4\alpha(R)\sigma^2(\Sigma_K)} \text{tr}\{S \nabla J(K)^\top \nabla J(K)S^\top\}
\]

\[
\leq \frac{\left\| \Sigma_K \right\| \left\| S \right\|^2}{4\alpha(R)\sigma^2(\Sigma_K)} \text{tr}\{\nabla J(K)^\top \nabla J(K)\}.
\]

Moreover, it can be observed that the gradient is orthogonal to the matrix space \( \mathcal{F} \). Define \( \Pi_F \) as the projection operator of a matrix onto \( \mathcal{F} \) and \( \Pi_{\perp F} \) onto its orthogonal space.

**Lemma 8.** Let \( K \in \mathcal{S} \). Then, we have \( \Pi_F(\nabla J(K)) = 0 \).

**Proof.** For any \( \Delta \in \mathcal{S} \), it holds that

\[
\text{tr}\{\Delta^\top \cdot \nabla J(K)\} = 2\text{tr}\{E_K \Sigma_K(S^\top)^\top \Delta^\top\} = 0.
\]

Hence, \( \nabla J(K) \) is orthogonal to \( \mathcal{F} \).

The above interesting fact implies that for an initial policy \( K^0 \), its projection \( \Pi_F(K^0) \) will not be affected by the gradient update (16). Along with Lemma 7, we have the following global convergence guarantees.

**Theorem 9.** For an appropriate stepsize \( \eta \) that is polynomial in problem parameters, e.g., \( \left\| A \right\|, \left\| B \right\|, \left\| S \right\|, \sigma(\Sigma_0), \sigma(Q), \sigma(R) \), the gradient update (16) converges to \( K^\infty = \Pi_F(K^0) + \Pi_{\perp F}(K^*) \) at a linear rate, i.e., for \( i \in \{0,1,\ldots\} \),

\[
J(K^{i+1}) - J^* \leq \left(1 - \frac{2\eta \alpha(R)}{\left\| \Sigma_K \right\| \left\| S \right\|^2} \right) (J(K^i) - J^*).
\]

Fig. 1. An illustration of \( K^0, K^\infty, K^* \) and \( \nabla J(K) \).

**Algorithm 1** The zero-order algorithm with two-point gradient estimate

**Input:** An initial policy \( K^0 \in \mathcal{S} \), the number of iterations \( N \), a smoothing radius \( r \), the stepsize \( \eta \).

1. for \( i = 0,1,\ldots,N - 1 \) do
2. Sample a perturbation matrix \( U^i \) uniformly from the unit sphere \( U^{mq-1} \).
3. Set \( K^1_i = K^i + r\sqrt{mq}U^i \) and \( K^2_i = K^i - r\sqrt{mq}U^i \).
4. Obtain \( \tilde{J}(K^1_i) \) and \( \tilde{J}(K^2_i) \) from (20).
5. Estimate the gradient

\[
\nabla J^i = \frac{1}{2\Delta}(\tilde{J}(K^1_i) - \tilde{J}(K^2_i))U^i.
\]
6. Update the policy by \( K^{i+1} = K^i - \eta\nabla J^i \).
7. end for

**Output:** A policy \( K^N \).

**Proof.** The convergence follows the same vein as the proof of (Fazel et al., 2018, Theorem 7).

To see why the resulting policy satisfies (17), we have by Lemma 8 that

\[
\Pi_F(K^\infty) = \Pi_F(K^0)
\]

and by Theorem 4 that

\[
\Pi_{\perp F}(K^\infty) = \Pi_{\perp F}(K^*)
\]

Fig. 1 illustrates the relation among \( K^0, K^\infty, K^* \) and \( \nabla J(K) \). Theorem 9 ensures that for any initial stabilizing policy \( K^0 \in \mathcal{S} \), the PG update in (16) converges to the solution set \( \mathcal{K} \) at a linear rate. In the following, we shall discuss its implementation when an explicit model \( (A, B, C) \) is unavailable.

**5. SAMPLE-BASED IMPLEMENTATION**

In this section, we propose a zero-order optimization algorithm to solve an optimal policy by only using input-output trajectories.

In the sample-based setting, the model \( (A, B, C) \) is unknown, and the gradient can only be estimated via zero-order information. However, it is challenging to evaluate the cost function as implementing \( u_0 = -Kz_{0,p} \) requires \( \{u_{-p}, y_{-p}, \ldots, u_{-1}, y_{-1}\} \) to be known.

To generate the required sequence, we use a random control policy \( u_t \sim \mathcal{N}(0, I) \) in \( t \in \{-p, \ldots, -1\} \). More specifically, we generate a trajectory by

\[
x_{-p} \sim \mathcal{N}(0, I) \quad \text{and} \quad u_t = \begin{cases} w_t, & t \in \{-p, \ldots, -1\} \\ -Kz_{t,p}, & t \in \{0, \ldots, T\} \end{cases}
\]

over, we show that the gradient is orthogonal to \( \mathcal{F} \), which reveals an important relation between the resulting solution and the initial policy.
Fig. 2. Convergence of the PG update in (16).

where \( w_t \) is an i.i.d. random sequence. By the dynamics in (1), \( x_0 \) satisfies

\[
x_0 = [c_p, A_p] [w_{0, p}, x_{-p}].
\]

Since \( C_p \) has full row rank, the distribution of \( x_0 \) generated by (18) satisfies Assumption 1. Accordingly, we replace the cost function in (2) by

\[
J(K) = E \left[ \sum_{t=0}^T (q_t^T Q y_t + u_t^T R u_t) \right]
\]

subject to (1), and (18).

Then, we estimate \( J(K) \) in (19) by one single sampled trajectory, i.e.,

\[
\tilde{J}(K) = \sum_{t=0}^T (q_t^T Q y_t + u_t^T R u_t)
\]

subject to (1), and (18) with the sampling time \( T \). In practice, \( T \) is often selected to be finite as the resulting approximation error of (20) decreases exponentially to zero w.r.t. \( T \) (Malik et al., 2019). We refer to (20) as warm-up cost evaluation since it involves another (random) policy in \( t = \{-p, \ldots, -1\} \).

We present our zero-order algorithm with warm-up cost evaluation in Algorithm 1. Particularly, we use a two-point method (Malik et al., 2019) to estimate the gradient in step 2-5. The parameter \( r \) is called the smoothing radius, used to control the variance of the gradient estimate. For the convergence analysis of Algorithm 1, one can apply standard results (Malik et al., 2019, Theorem 1) of zero-order methods.

6. SIMULATION

In this section, we validate our theoretical results via simulations. Moreover, we compare the performance of our IOF controller from Algorithm 1 with the SOF controller. The simulation is carried out using MATLAB 2021b on a laptop with 2.8GHz CPU. The code is provided in https://github.com/fuxy16/input-output-Feedback.

6.1 Example

We randomly generate a dynamical model \((A, B, C)\) with \( n = 4, m = d = 2 \) as

\[
A = \begin{bmatrix}
0.568 & 0.215 & 0.122 & -0.156 \\
-0.074 & -0.021 & -0.114 & -0.307 \\
0.568 & 0.211 & 0.047 & -0.604 \\
-0.455 & 1.141 & -0.204 & -0.478
\end{bmatrix}
\]

The simulation is carried out using MATLAB 2021b on a laptop with 2.8GHz CPU. The code is provided in https://github.com/fuxy16/input-output-Feedback.

6.2 Convergence of our PG method

In the model-based setting, we assume that \((A, B, C)\) is known, and perform (16) to validate the global convergence result in Theorem 9. Let the stepsize be \( \eta = 10^{-3} \). The initial policy \( K^0 \) is selected by first generating a random matrix with its elements being Gaussian and then normalizing it such that \( p(A - BK^0) = 0.8 \). Then, we conduct (16) where the gradient is computed using \((A, B, C)\) and display the optimality gap in the cost function in Fig. 2. The bold centreline denotes the mean of 20 independent trials and the shaded region demonstrates their variance. As expected from Theorem 9, the gap diminishes fast at a linear rate, and the randomness of \( K^0 \) only induces a small variance.

In the sample-based setting, we conduct Algorithm 1 to demonstrate the performance of our zero-order method. Let the sampling time in (20) be \( T = 20 \), the stepsize \( \eta = 10^{-5} \), the smoothing radius \( r = 0.2 \) and \( K^0 = 0 \). The convergence is shown in Fig. 3, where the variance originates from the warm-up process of (20).

6.3 Comparison with SOF

To show the merits of our new parameterization, we compare it with the SOF controller in the form of

\[
u_t = -K_s g_t,
\]

where \( K_s \) is solved by PG methods in Duan et al. (2022b).

Particularly, we consider two cases, \( d = 2 \) and \( d = 4 \). When \( d = 2 \), the matrix \( C \) is rank deficient and \( K_s \) is only guaranteed to be locally minimal (Duan et al., 2022b; Polyak, 1963). When \( d = 4 \), we sample a new model

Fig. 3. Convergence of Algorithm 1.

Table 1. The region of attraction radius \( r_{oa} \).

| Method                | Model-based (d=2) | Sample-based (d=2) | Model-based (d=4) | Sample-based (d=4) |
|-----------------------|-------------------|--------------------|-------------------|--------------------|
|                       |                   |                    |                   |                    |
| Optimal               | 8.282             | 10.562             | 19.649            | 26.112             |
| IOF                   | 8.987             | 11.438             | 19.793            | 26.428             |
| SOF                   | 13.013            | 15.519             | 20.611            | 29.467             |

This system is open-loop stable, and is both controllable and observable with \( p = 2 \). Let \( Q = I_4 \) and \( R = 0.01 I_2 \). Then, Assumption 2 is satisfied. In the following, we search over the matrix space \( \mathbb{R}^{2 \times 8} \) to find an optimal solution.
(A, B, C) randomly where the matrix C is invertible and hence $K_0$ is expected to have the same performance as the LQR controller (Duan et al., 2022b, Theorem 1). We set the stepsize by grid search which is $\eta = 10^{-5}$ for both IOF and SOF, and set other parameters as before. We conduct $10^5$ iterations of PG updates and compare the performance between the resulting IOF and SOF controllers. Their average costs of 20 independent trials are displayed in Table 6.2. In the case $d = 2$, the results are reasonable as the SOF only converges to local minima. Surprisingly, even in the case $d = 4$ the IOF controller still yields a lower cost. We note that the matrix space of the SOF problem is $\mathbb{R}^{2 \times 4}$, which is $\mathbb{R}^{2 \times 12}$ for the IOF problem. Thus, this result means that our PG method converges faster even though its gain matrix has a higher dimension.

7. CONCLUSION

In this paper, we have proposed a new parameterization of the policy, under which the PG method has been shown to globally converge to the solution set. We have also found some interesting properties such as the orthogonality of the gradient, and the invariance of the solution set to similarity transformation.

We now discuss some possible future works. Since this paper only considers the vanilla gradient descent method, it would be interesting to investigate the performance of both natural gradient and Gauss-Newton methods, which have been shown to have a faster convergence rate in the LQR problem (Fazel et al., 2018). It is also interesting to see whether the convergence can be preserved in the presence of process and measurement noises.

REFERENCES

Andrychowicz, O.M., Baker, B., Chociej, M., Jozefowicz, R., McGrew, B., Pachocki, J., Petron, A., Plappert, M., Powell, G., Ray, A., et al. (2020). Learning dexterous in-hand manipulation. The International Journal of Robotics Research, 39(1), 3–20.

Bertsekas, D. (2012). Dynamic programming and optimal control, volume 1. Athena Scientific, Massachusetts.

Duan, J., Cao, W., Zheng, Y., and Zhao, L. (2022a). On the optimization landscape of dynamical output feedback linear quadratic control. arXiv preprint arXiv:2201.09598.

Duan, J., Li, J., Li, S.E., and Zhao, L. (2022b). Optimization landscape of gradient descent for discrete-time static output feedback. In 2022 American Control Conference (ACC), 2932–2937. IEEE.

Fatkhullin, I. and Polyak, B. (2021). Optimizing static linear feedback: Gradient method. SIAM Journal on Control and Optimization, 59(5), 3887–3911.

Fazel, M., Ge, R., Kakade, S., and Mesbahi, M. (2018). Global convergence of policy gradient methods for the linear quadratic regulator. In International Conference on Machine Learning, 1467–1476.

Gravell, B., Esfahani, P.M., and Summers, T. (2020). Learning optimal controllers for linear systems with multiplicative noise via policy gradient. IEEE Transactions on Automatic Control, 66(11), 5283–5298.

Hu, B., Zhang, K., Li, N., Mesbahi, M., Fazel, M., and Baar, T. (2022). Towards a theoretical foundation of policy optimization for learning control policies. arXiv preprint arXiv:2210.04810.

Levine, S., Finn, C., Darrell, T., and Abbeel, P. (2016). End-to-end training of deep visuomotor policies. The Journal of Machine Learning Research, 17(1), 1334–1373.

Li, Y., Tang, Y., Zhang, R., and Li, N. (2021). Distributed reinforcement learning for decentralized linear quadratic control: A derivative-free policy optimization approach. IEEE Transactions on Automatic Control, 1–16. doi: 10.1109/TAC.2021.3128592.

Malik, D., Pananjady, A., Bhatia, K., Khamaru, K., Bartlett, P., and Wainwright, M. (2019). Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. In The 22nd International Conference on Artificial Intelligence and Statistics, 2916–2925.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529–533.

Polyak, B.T. (1963). Gradient methods for the minimization of functionals. USSR Computational Mathematics and Mathematical Physics, 3(4), 864–878.

Recht, B. (2019). A tour of reinforcement learning: The view from continuous control. Annual Review of Control, Robotics, and Autonomous Systems, 2, 253–279.

Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. Nature, 529(7587), 484–489.

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., and Abbeel, P. (2017). Domain randomization for transferring deep neural networks from simulation to the real world. In IEEE/RSJ international conference on intelligent robots and systems (IROS), 23–30.

Zhang, K., Hu, B., and Basar, T. (2021). Policy optimization for $H_2$ linear control with $H_\infty$ robustness guarantee: Implicit regularization and global convergence. SIAM Journal on Control and Optimization, 59(6), 4081–4109.

Zhao, F., Fu, X., and You, K. (2022). On the sample complexity of stabilizing linear systems via policy gradient methods. arXiv preprint arXiv:2205.14335.

Zhao, F., You, K., and Basar, T. (2021). Global convergence of policy gradient primal-dual methods for risk-constrained lqrs. arXiv preprint arXiv:2104.01901.

Zheng, Y., Sun, Y., Fazel, M., and Li, N. (2022). Escaping high-order saddles in policy optimization for linear quadratic gaussian (lqg) control. arXiv preprint arXiv:2204.00912.

Zheng, Y., Tang, Y., and Li, N. (2021). Analysis of the optimization landscape of linear quadratic gaussian (LQG) control. arXiv preprint arXiv:2102.04393.