Structured Output Feedback Control for Linear Quadratic Regulator Using Policy Gradient Method

Shokichi Takakura and Kazuhiro Sato, Member, IEEE

Abstract—In this article, we consider the static output feedback control for linear quadratic regulator problems with structured constraints under the assumption that system parameters are unknown. To solve the problem in the model free setting, we propose the policy gradient algorithm based on the gradient projection method and show its global convergence to $\varepsilon$-stationary points. In addition, we introduce a variance reduction technique and show both theoretically and numerically that it significantly reduces the variance in the gradient estimation. We also show in the numerical experiments that the model free approach efficiently solves the problem.

Index Terms—Data-driven control, gradient descent, linear quadratic regulator (LQR), model free control, nonconvex optimization, reinforcement learning (RL).

I. INTRODUCTION

Linear quadratic regulator (LQR), which is a well-studied framework in the optimal control theory, has been revisited from the reinforcement learning (RL) perspective. For policy gradient methods, the global linear convergence to the global optima was obtained in [1], [2]. To obtain structured policy, structured policy iteration of state feedback gains for LQR problems with a regularization term was proposed in [3] and the local linear convergence to a stationary point was provided. In addition, the projected gradient method for model-free state feedback LQR problems with convex constraints were studied in [4]. For the model-based setting, the projected method was studied in [5] and [6] considered linearly constrained problem for state feedback LQR problems.

However, it is difficult to observe the entire state. That is, only some outputs are available in practice. The static output feedback control is a practical approach to deal with such situations. For model-based control design, some iterative methods are found in [7] and recently, the global convergence of the gradient descent for output feedback LQR problems was shown in [8] using smoothness and Lipschitz continuity on the level sets of the LQR objective function. A model free algorithm was also proposed in [9] based on integral RL. However, policy gradient methods for static output feedback problems in the model free setting have not been well studied.

In this study, we consider a policy gradient method for the LQR problem with structured constraints for the static output feedback control under the assumption that system parameters are unknown, in contrast to many existing works [1], [2], [4], which studied the policy gradient method for state feedback LQR problems and [8], which studied gradient methods in the model-based setting. The structured constraints are naturally introduced due to the system structure, such as linear port-Hamiltonian systems [10].

Our Contribution: The main contributions of this article are summarized as follows.

1) To solve the LQR problem with structured constraints in the model free setting, we propose a policy gradient projection algorithm with a gradient estimation procedure.

2) We show the global convergence to $\varepsilon$-stationary points of our proposed algorithm using the LQR objective function properties, such as bounded sublevel sets, $L$-smoothness on sublevel sets, and dependency on horizon time. In addition, we show that the feedback gain obtained by the proposed method asymptotically stabilizes the closed-loop system. We also provide the sample complexity of the gradient estimation procedure.

3) We propose a variance reduction method using the baseline technique and show its suboptimality.

Article Organization: In Section II, we introduce the LQR problem for output feedback control with structured constraints. In Section III, we show some properties of the objective function on sublevel sets. In Section IV, we propose the gradient estimation method and the policy gradient projection algorithm in the model free setting. We then show that the algorithm outputs an $\varepsilon$-stationary point with high probability. In addition, we provide a variance reduction method and show its asymptotic optimality. In Section V, we conduct some numerical experiments and show properties of our proposed method. Finally, Section VI concludes this article.

Notation: For a vector $v \in \mathbb{C}^n$, $v^\top$ and $v'$ denote the transpose and conjugate transpose of $v$, respectively. The symbol $I$ and $O$ denote the identity matrix and the zero matrix, respectively. The symbol $S^n$ denotes the set of $n \times n$ symmetric matrices. For a matrix $A \in \mathbb{R}^{n \times m}$, $||A||$ and $||A||_2$ represent Frobenius and spectral norm of $A$, respectively. $\lambda_i(A)$ denotes the $i$th eigenvalue of $A$ indexed as $\text{Re}(\lambda_1(A)) \leq \cdots \leq \text{Re}(\lambda_n(A))$, and $\text{vec}(A) \in \mathbb{R}^{nm}$ denotes the vectorized form of $A$. For matrices $A, B \in \mathbb{R}^{n \times m}$, the inner product $(A, B)$ is defined as $(A, B) = \text{tr}(AB^\top)$ and $A \circ B$ denotes the Hadamard product of $A$ and $B$. Given a symmetric matrix $S \in S^n$, $\lambda_{\min}(S)(\lambda_{\max}(S))$ denotes the minimum (maximum) eigenvalue of $S$. Given a random variable $X$ which follows the distribution $D$, $E_X\{v[X]\}$ or just $E[X]$ denotes the expectation over $X \sim D$ and $V[X]$ denotes the variance of $X$. For $z \in \mathbb{C}$, $\text{Re}(z)(\text{Im}(z))$ denotes the real (imaginary) part of $z$.

II. PROBLEM FORMULATION

We consider the linear time-invariant (LTI) system

$$\dot{x}(t) = Ax(t) + Bu(t), \quad y(t) = Cx(t), \quad x(0) \sim D, \quad (1)$$

"
where \( x(t) \in \mathbb{R}^n \) is state, \( u(t) \in \mathbb{R}^m \) is input, \( y(t) \in \mathbb{R}^p \) is output, \( A \in \mathbb{R}^{n \times n} \), \( B \in \mathbb{R}^{n \times m} \), and \( C \in \mathbb{R}^{p \times n} \) are constant matrices, and \( D \) is a probability distribution over \( \mathbb{R}^n \). In this article, we assume that \( B \) and \( C \) are not zero matrices, and \( (A, B, C, D) \) is unknown unlike the situation in [8]. The infinite-horizon continuous-time LQR problem is formulated as

\[
\text{minimize } E_x(0) - D \left[ \int_0^\infty \left( y^\top(t)Qy(t) + u^\top(t)Ru(t) \right) dt \right]
\]

subject to (1)

with constant positive definite matrices \( Q \in \mathbb{R}^{p \times p} \) and \( R \in \mathbb{R}^{m \times m} \). The expectation is taken with respect to the initial state \( x(0) \sim D \). For the static output feedback \( u(t) = -Ky(t) \) with \( K \in \mathbb{R}^{m \times p} \) to system (1), the objective function (2) becomes \( f(K) := E_x(0) - D \int f(K; y(t)) dt \), where

\[
f(K; v) := \int_0^\infty \left[ v^\top e^{A_k t} C^\top (Q + K^\top RK) Ce^{A_k t} v \right] dt
\]

for \( v \in \mathbb{C}^n \). Then, the closed-loop is given by

\[
\dot{x}(t) = A_K x(t), \quad y(t) = C x(t),
\]

where

\[ A_K := A - BK. \]

In this article, we consider the constraints \( K \in \Omega \), where \( \Omega \subset \mathbb{R}^{m \times p} \) is a closed convex set that specifies the structural information of feedback gains. This is because a structured policy is often used in practical situations as follows.

1) \textbf{Decentralized Control:} In decentralized control, some components of \( K \) need to be 0 [11]. This implies that \( \Omega \) should be a certain linear subspace of \( \mathbb{R}^{m \times p} \).

2) \textbf{Linear Port-Hamiltonian System:} For a linear port-Hamiltonian system [10], if the feedback gain is positive semidefinite, the closed loop system is also a port-Hamiltonian system and passive. To ensure passivity, \( \Omega \) should be defined as the set of positive semidefinite matrices, which is closed and convex.

By using Bellman lemma [12], the problem (2) with structured constraints can be formulated as

\[
\text{minimize } f(K) = \text{tr}(X \Sigma) \text{ subject to } K \in \Omega \text{ and } A_K \text{ is Hurwitz},
\]

where

\[ \Sigma := E[x(0)x^\top(0)] \]

and \( X \) is the solution to

\[ A_K^X X + XA_K + C^\top (Q + K^\top RK + Q) C = 0. \]

It is difficult to solve (5), since \( f(K) \) is nonconvex and saddle points may exist [8]. Moreover, the feasible set may have exponentially many disconnected components [13]. Although an iterative method was proposed in [9] to obtain a suboptimal static output feedback gain in the model-free setting, it cannot be applied directly to problem (5) due to the constraint \( K \in \Omega \).

To develop a model-free algorithm with theoretical guarantees for solving problem (5), we impose the following assumption throughout this article.

\textbf{Assumption 1:}

1) \( \Sigma > 0 \).

2) The pair \((A, C)\) is observable.

3) There exists \( K_0 \in \Omega \) such that \( A_{K_0} \) is Hurwitz and \( K_0 \) is known. Since \( A_{K_0} \) is Hurwitz, there exist positive definite matrices \( G, H \) and a skew-adjoint matrix \( J \) such that \( A_{K_0} = (J - G)H \). The proof is found in [14]. Let \( H = L^\top L \) be the Cholesky decomposition. Using the coordinate transformation \( x'(t) = Lx(t) \), the closed-loop system (3) becomes

\[ \dot{x}'(t) = A_{K_0}'x'(t), \quad y(t) = C'x'(t), \]

where \( A_{K_0}' = LJJ^\top - LGL^\top, C' = CL^\top \). Since \( LJJ^\top \) and \( -LGL^\top \) are not zero matrices, and \( (A, B, C, D) \) is unknown unlike the situation in [8]. The infinite-horizon continuous-time LQR problem becomes

\[ E_x(0) - D \left[ \int_0^\infty \left( y^\top(t)Qy(t) + u^\top(t)Ru(t) \right) dt \right]
\]

subject to (1)

with constant positive definite matrices \( Q \in \mathbb{R}^{p \times p} \) and \( R \in \mathbb{R}^{m \times m} \). The expectation is taken with respect to the initial state \( x(0) \sim D \). For the static output feedback \( u(t) = -Ky(t) \) with \( K \in \mathbb{R}^{m \times p} \) to system (1), the objective function (2) becomes \( f(K) := E_x(0) - D \int f(K; y(t)) dt \), where

\[
f(K; v) := \int_0^\infty \left[ v^\top e^{A_k t} C^\top (Q + K^\top RK) Ce^{A_k t} v \right] dt
\]

for \( v \in \mathbb{C}^n \). Then, the closed-loop is given by

\[ \dot{x}(t) = A_K x(t), \quad y(t) = C x(t), \]

where

\[ A_K := A - BK. \]

If \( K \notin S \), there exists an eigenvalue \( \mu \) of \( A \) such that \( \text{Re}(\mu) > 0 \) and \( f(K) \) goes to infinity.

\textbf{Remark 1:} The objective function of problem (2) is not a standard LQR cost, as in some previous researches [15, 16]. While similar convergence properties to the standard LQR cost can be obtained for our formulation in the model-based setting if \( (A, C) \) is observable [8], more detailed studies of the objective function properties are necessary for the model-free version of the convergence analysis.

\section{III. PROPERTIES OF THE OBJECTIVE FUNCTION}

In this section, we prove some properties of the objective function \( f(K) \) in (5) for the convergence analysis of the gradient method presented in Section IV.

\subsection{A. Norm Bounds}

In this section, we show some matrix norm inequalities. We define the sublevel set by \( S(a) = \{ K \in S \mid f(K) \leq a \} \), where \( S \) is defined as (6). Thus, all elements in \( S(a) \) are stabilizing feedback gains.

Using the same argument as [8, Lemma C.2], we have \( \| K \| \leq \kappa(a) \) for \( K \in S(a) \), where

\[
\kappa(a) := \frac{2\|B\|_2\|C\|_2a}{\lambda_{\min}(\Sigma)\lambda_{\min}(R)\lambda_{\min}(CC^\top)} + \|A\|_2^2 \|B\|_2\|C\|_2^2.
\]

Next, for \( K \in S(a) \), we provide an upper bound on the solutions to the following Lyapunov equations:

\[
A_K^X X + XA_K + C^\top (Q + K^\top RK) C = 0, \]

\[
A_K Y + YA_K^\top + \Sigma = 0, \]

\[
A_K Y' + Y'A_K^\top - (\text{BECV} + (\text{BECV})^\top) = 0,
\]

where \( A_K \) is defined in (4) and \( E \in \mathbb{R}^{m \times p} \) is a given matrix. Note that \( X, Y, \) and \( Y' \) uniquely exist, because \( K \in S(a) \) implies that \( A_K \) is Hurwitz [12]. To simplify the notation, using \( \sigma := -\frac{1}{2}\lambda_{\max}(A_{K_0} + A_{K_0}') > 0 \), we define

\[
\xi := \frac{1}{4\|B\|_2\kappa(a)}, \quad \chi(a) := \frac{a}{\lambda_{\min}(\Sigma)},
\]

\[
\Theta(a) := \max \left( \frac{\xi^2\lambda_{\min}(Q)}{\sigma}, \frac{\|\Sigma\|}{\sigma} \right),
\]

\[
\Theta'(a) := 2\|B\|_2\|C\|_2\Theta(a)^2 \frac{\lambda_{\min}(\Sigma)}{\lambda_{\min}(\Sigma)},
\]

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
where $K_0$ satisfies 3 in Assumption 1.

**Lemma 1:** Let $X, Y, Y'$ be the solution to (7), (8), and (9), respectively. Assume that $\|X\|_F = 1$. Then, for any $K \in S(a)$, $\|X\|_2 \leq X(a)$, $\|Y\|_2 \leq Y(a)$, $\|Y'\|_2 \leq Y'(a)$.

**Proof:** See Appendix A.

**B. L-Smoothness of $f(K)$**

A differentiable function is called $L$-smooth if its gradient is $L$-Lipschitz continuous. For our objective function, we have the following result.

**Theorem 1:** For any $a \in \mathbb{R}$, $f(K)$ in (5) is $L$-smooth on $S(a)$ with the constant

$$L := 2\lambda_{\text{max}}(R) \|C\|_F^2 \|x\|_F \|
+ \left(\sqrt{n} \|R\|_F \|\alpha\| \|C\|_F + n \|B\|_F \| \alpha \| \right) \|y\| \|\|C\|_F \|,$$

where $y(\alpha)$ and $y'(\alpha)$ are defined in (10) and (11).

**Proof:** Theorem 3.15 in [8] cannot be applied to our setting directly, because $\lambda_i(C'QC)$ may be 0. However, by replacing the norm bounds in [8, proof of Theorem 3.15] with those in Lemma 1, we obtain the result.

**IV. MODEL FREE ALGORITHM**

In this section, we consider problem (5) in the model free setting. That is, we assume that $(A, B, C, D)$ in system (1) is unknown. First, we introduce a gradient estimation algorithm based on the derivative-free optimization approach [1], and we propose algorithm 2 to calculate the stochastic estimate $\hat{f}(K)$ of the gradient $\nabla f(K)$.

**Algorithm 1:** Gradient Estimation.

**Require:** $K \in \Omega$, $N > 0$, $\tau > 0$, $\tau > 0$

1: for $i = 1$ to $N$ do
2: Sample $U_i$ from the uniform distribution $S$ over matrices with $\|U_i\|_F = \sqrt{mp}$.
3: Simulate the system
   
   $\dot{x}_i(t) = Ax_i(t) + Bu_i(t)$,  $y_i(t) = Cx_i(t)$,
   
   where $u_i(t) = -(K + rU_i)y_i(t)$, $x_i(0) \sim D$ until time $\tau$ and calculate the empirical cost
   
   $c_i = \int_0^\tau [y_i(t)'Qy_i(t) + u_i(t)'^TRu_i(t)]dt$.
4: end for
5: Define the estimated gradient by

$$\hat{\nabla} f(K) = \frac{1}{rN} \sum_{i=1}^N c_i U_i.$$  (13)

6: return $\hat{\nabla} f(K)$.

In this section, we assume the following in addition to Assumption 1.

**Assumption 2:**

1) $U_i$ and $x_i(0)$ are independent.
2) $K \in S(a)$ with a given constant $a$.
3) The distribution $D$ of initial state $x(0) \sim D$ satisfies $\|x(0)\|_2 \leq P$ a.s. for a constant $P > 0$.

The second assumption is justified in Section IV-B. The following lemma ensures $K + rU \in S(2a)$ for sufficiently small $r$.

**Lemma 2:** There exists $r_0 > 0$ such that for any $r \leq r_0$, $K \in S(a)$ and $U$ such that $\|U\| = \sqrt{mp}$, we have $K + rU \in S(2a)$.

**Proof:** Using the same argument as [2, Lemma 4], we obtain the result.

The following theorem is an extension of [1, Lemma 27 in Supplementary material] to output feedback control. To prove this, we derive some inequalities in Appendix B.

**Theorem 2:** Let $\nabla f(K)$ be defined as (13). For any $\epsilon > 0$ and $\delta > 0$, set $r = O(\epsilon)$, $\tau = O(\log(1/\epsilon))$, $N = O((\log(1/\delta)/\epsilon^2))$. Then, we have

$$\left\|\nabla f(K) - \nabla f(K)\right\|_F \leq \epsilon$$

(14)

for any $K \in S(a)$, with probability greater than $1 - \delta$.

**Proof:** See Appendix B.

**B. Convergence Properties**

In this section, we prove the global convergence of the policy gradient method in the model free setting. We show a model free control algorithm, policy gradient projection, in Algorithm 2. The positive integer $T$ is the iteration number, $\alpha$ is the step size, $K_0$ is the initial point of the feedback gain $K$, and proj is the projection onto $\Omega$ with respect to the Frobenius norm. The termination condition $\|K_{i+1} - K_i\|_F \leq \epsilon\alpha$ is added for technical reasons.

**Lemma 3:** describes the property of orthogonal projections, which is essential to our convergence analysis.

**Lemma 3:** Let $\text{proj} : \mathbb{R}^{mp} \rightarrow \Omega$ be an orthogonal projection onto $\Omega$. For any $x \in \Omega$ and $y \in \mathbb{R}^{mp}$, we have $\langle x - \text{proj}(y), y - \text{proj}(y) \rangle \leq 0$.

**Proof:** See [17, Th. 6.41].

The following definition is required to show our main result.

**Definition 1:** For positive constants $\alpha$ and $\epsilon$, $K$ is called an $\epsilon$-stationary point if $\|G_\alpha(K)\|_F \leq \epsilon$, where $K^+ := \text{proj}(K - \alpha\nabla^2 f(K))$ and $G_\alpha(K) := \frac{1}{\alpha}(K^+ - K)$.

The following theorem is a main result, which is an extension of [8, Th. 4.2] for the model-free and constrained problems. The proof is based on the proof for projected gradient method without gradient error for $L$-smooth functions [17]. However, in the presence of gradient errors,
the termination condition and some extra arguments to bound the effect of the difference between the true and estimated gradients are required.

Theorem 3: Assume that the constants $N, r, \tau$ satisfy the condition in Theorem 2 with $\varepsilon' = \lambda\varepsilon$ for the given constants $0 < \lambda < 1, \varepsilon > 0, \delta > 0$. Let $\{K_i\}_{i=0}^T$ be the sequence generated by the Algorithm 2, where $T'$ is the total number of iterations, which can be different from $T$ due to the terminate condition. For a step size $\alpha \in (0, \frac{2(1-\lambda)}{L})$, where $L$ is the Lipschitz constant of $\nabla f$ on $S_0$, we have the following result with probability greater than $1 - \delta$.  

1) The sequence $\{K_i\}_{i=0}^T$ remains in $S$ and $\{f(K_i)\}_{i=0}^T$ is strictly decreasing. That is, for any $0 \leq i \leq T' - 1$

$$f(K_{i+1}) < f(K_i).$$  

2) If $T > \frac{f(K_0)}{\varepsilon^2 \alpha^2 \left(\frac{1}{2} - \frac{1-\lambda}{\alpha}\right)}$, $K_T$ is a $(1 + \lambda)\varepsilon$-stationary point.

Proof: We define $G_{\alpha}(K) = \frac{1}{\alpha} v$, where $v := \dot{K}^+ - K$ and $\dot{K}^+ := \text{proj}(K - \alpha \nabla f(K))$. First, we show that if $\|G_{\alpha}(K)\|_p \leq \varepsilon$ and (14) holds, $K \in S_0$ is a $(1 + \lambda)\varepsilon$-stationary point.

$$\|G_{\alpha}(K)\|_p \leq \|G_{\alpha}(K)\|_p \leq \frac{1}{\alpha} \|K^+ - \dot{K}^+\|_p \leq \varepsilon + \|\nabla f(K) - \nabla f(K^+)\|_p \leq (1 + \lambda)\varepsilon.$$  

The second inequality holds, because projections onto convex sets are contractive, and the last inequality follows from (14).

Next, we show $\|G_{\alpha}(K)\|_p \leq \varepsilon$ with high probability. The termination condition ensures $\|G_{\alpha}(K)\|_p \geq \varepsilon$ for $i = 0, \ldots, T' - 1$. Assume that $\|G_{\alpha}(K)\|_p \geq \varepsilon$ and (14) with $\varepsilon' = \lambda\varepsilon$ holds for $K \in S_0$. We define $K_t := K + tv$ and $t' := \max\{t > 0 : f(K_t) \leq f(K_0), 0 \leq t' \leq t\}$. Then, Lemma 3 yields

$$\langle \nabla f(K), v \rangle \leq -\frac{1}{\alpha} \|v\|_p^2$$  

and $L$-smoothness of $f(K)$ on $S_0$ implies

$$f(K_t) - f(K) \leq \langle \nabla f(K), K_t - K + \frac{L}{2} t^2 \|K_t - K\|_p^2 \rangle \leq t\langle \nabla f(K), v \rangle + \frac{Lt^2}{2} \|v\|_p^2.$$  

By adding both sides of (16) multiplied by $t$ and (17), we obtain

$$f(K_t) - f(K) \leq \left(\frac{Lt^2}{2} - t - \frac{1-\lambda}{\alpha}\right) \|v\|_p^2,$$  

where we used (14) and $G_{\alpha}(K) = \|v\|/\alpha > \varepsilon$. For $t = t'$, we have

$$0 = f(K_{t'}) - f(K) \leq \left(\frac{Lt'}{2} - t' - \frac{1-\lambda}{\alpha}\right) \|v\|_p^2.$$  

Since $\|v\|_p > 0$, we have $t' \geq \frac{2(1-\lambda)}{L\alpha} \geq 0$. Therefore, (18) holds for $t = t'$ with $K = K_t$ and $t = t'$ leads us to

$$f(K_{t+1}) - f(K_t) \leq \left(\frac{L}{2} - \frac{1-\lambda}{\alpha}\right) \|K_{t+1} - K_t\|_p^2 < 0,$$  

because $\alpha \in (0, \frac{2(1-\lambda)}{L})$. Thus, (15) holds for any $0 \leq i \leq T' - 1$. If $T' < T$, the termination condition ensures $\|G_{\alpha}(K_{T'})\|_p \leq \varepsilon$. Therefore, it suffices to show $T' < T$. Since $T' \leq T$ by definition, suppose that $T' = T$. Then, we obtain

$$f(K_0) \geq f(K_0) - f(K_T) \geq \left(\frac{1-\lambda}{\alpha} - \frac{1}{2}\right) T\varepsilon^2 \alpha^2.$$  

The assumption $T > \frac{f(K_0)}{\varepsilon^2 \alpha^2 \left(\frac{1}{2} - \frac{1-\lambda}{\alpha}\right)}$ yields $f(K_0) > f(K_T)$, which is a contradiction. Thus, $T' < T$, that is, $\|G_{\alpha}(K_{T'})\|_p \leq \varepsilon$. From Theorem 2, the probability that (14) holds for $K = K_t (i = 0, \ldots, T')$ is greater than $1 - \delta(T')^{-1} \geq 1 - T\delta$. This completes the proof.  

The convergence rate $T = O(1/\varepsilon^2)$ is essentially the same as the rate of the projected gradient method without gradient error for $L$-smooth functions [17]. For sample complexity, the total number of samples $T'N = O(\log(1/\delta)/\varepsilon^2)$ is worse than $O(\log(1/\delta)/\varepsilon^2)$ of the zeroth-order proximal gradient descent with two points evaluation in [18], since we cannot evaluate the two cost function values with two different feedback gains for the same initial state due to the randomness of the initial state. Note that the same rate to ours was obtained for discrete-state feedback LQR problems in the model free setting [4] but not known for the model free and output feedback setting.

C. Variance Reduction

Policy gradient methods tend to suffer from a large variance, which leads to slow learning [19]. The use of baseline is one of the variance reduction techniques for policy gradient methods [19]. State-dependent functions are often used as a baseline, because it does not add any bias to the estimated gradient [20], [21]. In this section, we propose to use the finite horizon cost function as a baseline and show its optimality.

For a baseline function $b(x)$, the estimated gradient $\nabla f(K)$ is defined as

$$\nabla f(K) := \frac{1}{rN} \sum_{i=1}^N \{\tilde{f}_i(K + ru_i; x_i(0)) - b(x_i(0))\} U_i,$$  

where the finite horizon cost function is defined as

$$\tilde{f}_i(K; x(0)) := \int_0^T y(t)(Q + K^R R) y(t) dt,$$  

which satisfies $\lim_{T \to \infty} \tilde{f}_i(K; x(0)) = f(K; x(0))$. Because

$$\nabla f(K) = \frac{1}{rN} \sum_{i=1}^N \tilde{f}_i(K + ru_i; x_i(0)) U_i,$$  

the estimated gradient $\nabla f(K)$ satisfies

$$E\left[\nabla f(K)\right] = E_{x_i(0) \sim D, u_i \sim S} \left[\nabla f(K)\right]$$  

$$- \frac{1}{rN} \sum_{i=1}^N E_{x_i(0) \sim D, u_i \sim S} \left[b(x_i(0)) U_i\right]$$  

$$= E_{x_i(0) \sim D, u_i \sim S} \left[\nabla f(K)\right].$$  

The second equality holds from the assumption that $U_i$ and $x_i(0)$ are independent, and $E[U_i] = 0$. Thus, the bias in $\nabla f$ is the same as the one in $\nabla f$.

In terms of the variance, the baseline $b(x(0)) = \tilde{f}_i(K; x(0))$ is almost optimal for small $r$.

Theorem 4: For $a > 0$, $r \leq r_0 > 0$, and $K \in S(a)$, the optimal baseline $b^*(x(0))$ which minimizes the variance of the estimated gradient (19) is given by

$$b^*(x(0)) = E_{U \sim S} \tilde{f}_i(K + ru; x(0)),$$  

where $S$ is defined in Algorithm 1 and

$$\lim_{r \to 0} b^*(x(0)) = \tilde{f}_i(K; x(0)).$$
such that and is independent of the choice of and in the same manner as be a linear mpE with probability 1.

In the model free setting, the initial state in the estimation procedure ofaky the estimation error is independent of the choice of a baseline as well, that is,

\[
E_x(\tilde{f}_r(K + ru; x(0)) - b(x(0)))U = E_x(\tilde{f}_r(K + ru; x(0)) - b(x(0)))U.
\]

Then, the second term in (23) is independent of the choice of a(x) and we just need to minimize the first term in (23)

\[
E_x(\tilde{f}_r(K + ru; x(0)) b(x(0)))U = mpE_x(\tilde{f}_r(K + ru; x(0)) b(x(0)))^2.
\]

Since the expectation minimizes the mean squared error for any (x), the optimal baseline is given by (21), Equation (22) follows from the continuity of \(\tilde{f}_r(K + ru; x(0))\).

Based on Theorem 4, we propose to use \(\tilde{f}_r(K; x(0))\) as a baseline. In the model free setting, \(\tilde{f}_r(K; x(0))\) cannot be computed directly for a given (x) in the same manner as \(c_i\), because we cannot specify the initial state in the estimation procedure of \(c_i\). Therefore, we provide the estimation procedure for \(\tilde{f}_r(K; x(0))\) in Algorithm 3. In the following, we define \(g(t; x(0)) = |y(t)| |g(t + h_1)| \cdots |y(t + h_{d-1})|\) in \(R^{d}\) for \(0 \leq h_1 < h_2 < < h_{d-1} = T\).

**Algorithm 3: Estimate \(\tilde{f}_r(K; x(0))\)**

**Require:** \(K \in S, x(0) \in R^n, s > 0\).

1. \(i = 1\) to \((n+1)/2\) do
2. Simulate system (1) for \(x(0) \sim D\) until time s.
3. end for
4. Solve the following equations for \(\hat{P}(X)\).

\[
g(0; x_i(0)) P(K) g(0; x_i(0)) = \tilde{f}_r(K; x_i(0)) + g(x_i(0)) P(K) g(x_i; x_i(0)), \quad (24)
\]

where \(\tilde{f}_r\) is defined as (20).

5. Define \(\tilde{f}_r(K; x(0)) = \tilde{g}(0; x(0)) P(K) \tilde{g}(0; x(0)) - \tilde{g}(t; x(0)) P(K) \tilde{g}(t; x(0)).\)
6. return \(\tilde{f}_r(K; x(0)).\)

The following theorem ensures that the estimated cost \(\tilde{f}_r(K; x(0))\) is equal to \(\tilde{f}_r(K; x(0))\).

**Theorem 5:** For any \(T > 0, D > 2(n+1) + \frac{T}{2} \beta, \) where \(\beta = 2(|A| + |B| + |C| |x_k(a), if x(0), x_i(0), x(0), x_i(0)) \geq x_j(s) x_j(s) \geq \frac{n+1}{2}\) and \(\frac{n+1}{2}\) are linearly independent on \(S^d\) for s in Algorithm 3, then \(\tilde{f}_r(K; x(0)) = \tilde{f}_r(K; x(0))\) for any \(K \in S\) and \(x(0).\)

**Proof:** See Appendix C.

The following theorem shows that the assumption of Theorem 5 holds with probability 1.

**Theorem 6:** If the distribution \(D\) has a probability density function, \(x(0) x_i(0) \sim f(x_i; x_i) (j = 1, \ldots, \frac{n+1}{2})\) is linearly independent on \(S^d\) with probability 1.

**Proof:** Let \(v_i = x(0) x_i(0) - x_i(s) x_i(s)\) and \(V_u\) be a linear subspace generated by \(\{v_i\}_{i=1}^m (1 \leq m \leq \frac{n+1}{2})\). For \(m < \frac{n+1}{2}\), \(V_m\) is a proper subspace of \(S^d\) and there exists \(v_m \in S^d\) orthogonal to \(V_m\). Since \(x_i(s) = e^{A^k s} x_i(0)\), we have

\[
\{v_{m+1}, v_m\} = \langle x_{m+1}(0) x_{m+1}(0) - e^{A_k s} x_{m+1}(0) x_{m+1}(0) e^{A_k s}, v_m \rangle
\]

\[
= x_{m+1}(0) v_{m+1}(0),
\]

where \(v_{m+1} = e^{A_k s} v_{m+1} e^{A_k s}\). Note \(v_m \neq 0\), because \(A_k\) is Hurwitz and the solution to the discrete Lyapunov equation \(v = e^{A_k s} v e^{A_k s}\) is only \(v = 0\). Since \(v_m \in S^d\), there exists an orthogonal matrix \(U\) such that \(U^t v_m = \text{diag}(\mu_1, \ldots, \mu_m)\), where \(\{\mu_i\}_{i=1}^m\) are eigenvalues of \(v_m\). Without loss of generality, we assume \(\mu_m \neq 0\), since \(v_m \neq 0\). For \(x(0), x_i(0) \sim D\), \(z = U x_i(0)\). Then, \(x_{m+1}(0) v_{m} x_{m+1}(0) = \sum_{i=1}^m \mu_i z_i^2\). Since \(D\) has a probability density function, the distribution of \(z\) has also a probability density function \(g(1, \ldots, z_n) > 0\), the conditional probability density function of \(z_n\) is given by \(g(z_{1}, \ldots, z_n) d z_n = 0\), and the conditional probability that \(z_n\) satisfies \(\sum_{i=1}^m \mu_i z_i^2 = 0\) is zero, because there are at most two \(z_n\) in \(\mathbb{R}\) which satisfy \(\sum_{i=1}^m \mu_i z_i^2 = 0\). Therefore, the probability that \(\{v_{m+1}, v_m\} = x_{m+1}(0) v_{m+1}(0) = 0\) is zero. That is, \(v_{m+1} \notin V_m\) with probability 1. By induction, we obtain the result.

**V. NUMERICAL EXPERIMENTS**

In this section, we numerically demonstrate that the policy gradient projection algorithm can solve the LQR problem efficiently in the model free setting. Based on [8], we consider the problem (5) with \(A = (J - G) H, B = \text{ones}(10, 4) + \frac{1}{2} \text{randn}(10, 4), C = \text{ones}(2, 10) + \frac{1}{2} \text{randn}(2, 10), Q = I, R = I,\) where \(J = J - J^T, J = \text{randn}(10, 10), G = G^T, G = \text{randn}(10, 10), H = H^H, H = \text{randn}(10, 10),\) and \(\text{randn}(a,b)\) is a \(a \times b\) matrix of ones, \(\text{randn}(a,b)\) is a \(a \times b\) matrix with all entries distributed as the uniform distribution on \([0, 1]\), and \(\text{randn}(a,b)\) is a \(a \times b\) matrix with all entries distributed as the standard normal distribution. We assume the distribution \(D\) is the uniform distribution on \([-1, 1]^n\). Since \(J\) is skew-adjoint and \(G, H\) are positive definite, \(A\) is Hurwitz, as mentioned in Section II. Therefore, \(A_{K_0}\) is Hurwitz for \(K_0 = 0\). We set the parameters \(\tau = 0.01, \tau = 100\) and define \(\Omega\) by

\[
\Omega = \{K \in R^{4 \times 2} | K o S = 0\}, \quad S = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}^T.
\]

Fig. 1 illustrates that the mean and standard deviation of 20 trials of the relative error \(\frac{|E_y(K) - V_f(K)|}{|E_y(K)|}\) in gradient estimation. The relative error with variance reduction was much smaller than that of Algorithm 1.

Fig. 2 illustrates the mean and variance of 20 trials of the convergence curve of Algorithm 2, where we set \(\alpha = 2 \times 10^{-4}, 2 \times 10^{-5}\) and
Fig. 1. Relative error in gradient estimation.

Fig. 2. Convergence curve of Algorithm 2.

N = 15 for Algorithm 2 with baseline and N = 70 for Algorithm 2 without baseline. This is because the estimation procedure of the baseline requires additional \( n_{\alpha} = 55 \) samples. Because of the large variance of the estimated gradient, Algorithm 2 without baseline made the system unstable quickly in the case of \( \alpha = 2 \cdot 10^{-4} \). As shown in Fig. 2, the convergence rate of Algorithm 2 is sublinear and Algorithm 2 with baseline is more stable than Algorithm 2 without baseline even if we take into account additional 55 samples required to estimate the baseline.

VI. CONCLUSION

In this article, we considered the nonconvex optimization problem with convex constraints based on the output feedback version of LQR problems under the assumption that system parameters are unknown. To solve the problem, we proposed the policy gradient algorithm based on the gradient projection method and the zero-order optimization. We proved its global convergence to \( \varepsilon \)-stationary points with high probability. We also proposed the variance reduction method using the baseline technique and proved that it is almost optimal. In the numerical experiments, we showed that the baseline technique significantly reduces the variance in the gradient estimation and the model free method can achieve low LQR cost.

Policy gradient projection can be extended to the objective function with regularization terms using the proximal gradient method. In this setting, we are able to consider tradeoffs between cost function and structure, such as sparsity [3]. However, the convergence analysis would be more difficult, and it is left for a future work. In addition, the convergence of the gradient method with fixed step size could be slow, since the smooth constant \( L \) can be large depending on the initial feedback gain \( K_0 \). To overcome this issue, the gradient method with adaptive step size in the model-based setting was considered in [5] and optimization methods on the Riemannian manifolds were studied in [6]. Therefore, applying the adaptive step size to the model free algorithm is one of the important directions of future works. Other interesting directions of future works would be analysis for the natural policy gradient method [22] or other variants of the policy gradient method.

APPENDIX

A. Proof of Lemma 1:

Proof: For \( X \), see [2, Lemma 16]. Let \( \mu \) be the largest eigenvalue of \( Y \) and \( v \) be a normalized eigenvector corresponding to \( \mu \). Note that \( \| Y \|_2 = \mu \) since \( Y \succeq 0 \). In the following, we consider the case \( \| C v \| \geq \xi \) and the case \( \| C v \| \leq \xi \) separately. First, we consider the case \( \| C v \| \geq \xi \). Using \( f(K) = \text{tr}(Y C(C+K^T K))C \), we have \( \alpha \geq \text{tr}(Y(C+C^T K^T K)) \geq \text{tr}(Y C^T C) \). Since \( \mu v v^T \succeq Y \), \( \text{tr}(Y C^T C) \geq \text{tr}(\mu v v^T C^T C) = \mu \text{tr}(C^T C) \). Therefore, \( \alpha \geq \mu \text{tr}(C^T C) \geq \mu \| C v \|^2 \min(Q) \geq \mu \xi^2 \lambda_{\min}(Q) \) and \( \mu \leq \frac{a}{\xi^2 \min(Q)} \). Next, we consider the case \( \| C y \| \leq \xi \).

B. Proof of Theorem 2:

The total error \( \| \hat{\nabla} f(K) - \nabla f(K) \|_F \) can be divided into the bias term \( \| E[\hat{\nabla} f(K)] - \nabla f(K) \|_F \) and the variance term \( \| \hat{\nabla} f(K) - E[\hat{\nabla} f(K)] \|_F \).

First, we bound the bias term. The estimated gradient \( \hat{\nabla} f(K) \) in (13) can be expressed in the form

\[
\hat{\nabla} f(K) = \frac{1}{N_T} \sum_{i=1}^{N_T} \hat{f}_i(K + r U_i; x_i(0)) U_i.
\]

For any initial state \( x(0) \) and \( r > 0 \), we define the smooth function \( g_r(K) \) by \( g_r(K) := E_{i \sim S}[f(K + r U)] \), where \( S \) is the uniform distribution over the set \( \{ U \in \mathbb{R}^{n_T \times p} \mid \| U \|_2 \leq \sqrt{mp} \} \). Then, the bias in \( \hat{\nabla} f(K) \) can be divided into two parts as follows:

\[
\left\| \nabla f(K) - E \left[ \hat{\nabla} f(K) \right] \right\|_F \leq \left\| \nabla f(K) - \nabla g_r(K) \right\|_F + \left\| \nabla g_r(K) - E \left[ \hat{\nabla} f(K) \right] \right\|_F,
\]

where the expectation is taken over \( x_i(0) \sim \mathcal{D} \) and \( U_i \sim S \).

For the first term in (27), we have the following bound.

\[
\text{Lemma 4: For any } K \in S(a) \text{ and } r \leq r_0, \| \nabla f(K) - \nabla g_r(K) \|_F \leq L r \sqrt{mp}, \text{ where } L \text{ is the Lipschitz constant defined as (12) of } \nabla f \text{ on } S(2a).\]

Proof: From \( L \)-smoothness of \( f \), we have \( \| \nabla f(K) - \nabla f(K + r U) \|_F \leq L r \sqrt{mp} \). Therefore, \( \| \nabla f(K) - \nabla g_r(K) \|_F = E[\| \nabla f(K) - \nabla f(K + r U) \|_F] \leq L r \sqrt{mp} \).
Lemma 26 in Supplementary material of [1] implies $\nabla g_t(K) = E[\frac{1}{2} f(K + rU; x_0)U]$, and (26) yields $E[\nabla f(K)] = E[\frac{1}{2} f_t(K + rU; x_0)U]$, where the expectation is taken over $x_0 \sim D$ and $U \sim S$. By using these relations, we have the following upper bound of the second term in (27):

$$
\left\| \nabla g_t(K) - E \left[ \nabla f(K) \right] \right\|_F \\
\leq \frac{1}{r^N} \sum_{i=1}^N E \left[ \left| \tilde{f}(K + rU; x_i) - \tilde{f}_t(K + rU; x_i) \right| \right] \left\| U_i \right\|_F.
$$

(28)

To bound the right-hand side, we introduce the following lemma.

**Lemma 5:** For $K \in S(a)$ with $a \in \mathbb{R}$ and $x(t)$, which follows (3), we have

$$
\|x(t)\|^2 \leq \frac{2Q(a)\|y(0)\|}{\lambda_{\min}(\Sigma)} e^{-\lambda_{\min}(\Sigma)/2} \|x(0)\|^2,
$$

where $\lambda(a) = \|A\|^2 + \|B\|^2/\|C\|^2_2 \kappa(a)$.

**Proof:** From [2, Lemma 12], we have $\|e^{At}\|^2 \leq (\|Y\|^2/\lambda_{\min}(Y)) e^{-\lambda_{\min}(\Sigma)/2}$. Therefore, $\|x(t)\|^2 \leq (\|Y\|^2/\lambda_{\min}(Y)) e^{-\lambda_{\min}(\Sigma)/2} \|x(0)\|^2$.

(30)

Lemma 1 yields $\|Y\|^2 \leq \|y(a, t)\|$. This and [8, Lemma A.5] imply

$$
\lambda_{\min}(Y) \geq \frac{\lambda_{\min}(\Sigma)}{2} \frac{\|A\|^2}{\|A\|^2 + \|B\|^2/\|C\|^2_2 \kappa(a)}.
$$

(31)

Substituting these inequalities into (30), we obtain (29).

**Lemma 6:** For any $r \geq 0$, $x \in S(a)$,

$$
\left\| \nabla g_t(K) - E \left[ \nabla f(K) \right] \right\|_F = O(e^{-\eta r}/r),
$$

where $\eta = \lambda_{\min}(\Sigma)/\|y(0)\|$. $\lambda_{\min}(\Sigma)$

**Proof:** Lemma 2 implies that $K + rU \in S(2a)$, and we have

$$
\left\| \nabla g_t(K) - E \left[ \nabla f(K) \right] \right\|_F \leq \frac{1}{r} E \left[ x(0)^\top x(0) \right] \left\| x(0) \right\|_F \\
\leq \sqrt{\eta r} \left\| x(0) \right\|_2 E \left[ \|x(0)\|^2 \right],
$$

where $x(t)$ follows $\dot{x}(t) = A_{K+rU} x(t)$, and we used the fact

$$
\tilde{f}(K + rU; x(0)) - \tilde{f}_t(K + rU; x(0)) = x(0)^\top x(0),
$$

where $X$ is the solution to (7). From Lemma 5, we have

$$
\|x(t)\|^2 \leq \frac{2Q(a)\|y(0)\|}{\lambda_{\min}(\Sigma)} e^{-\lambda_{\min}(\Sigma)/2} \|x(0)\|^2.
$$

Thus, (31) holds, because $E[\|x(0)\|^2] = \text{tr}(\Sigma)$. Next, we obtain an upper bound of the variance term.

**Lemma 7:** For any $r > 0$, $x \in S(a)$, if $N = O((\log 1/\delta)/\epsilon^2)$, we have

$$
\Pr \left( \left\| \nabla f(K) - E \left[ \nabla f(K) \right] \right\|_F \geq \epsilon \right) \leq \delta.
$$

**Proof:** Using matrix Bernstein inequality [23], we obtain the result in the same way with [1, Lemma 27 in Supplementary material].

Combining Lemma 4, 6, and 7 completes the proof of Theorem 2.

**C. Proof of Theorem 5:**

For any $t \geq 0$, the observation $\tilde{y}(t; x(0))$ is determined by $\tilde{y}(t; x(0)) = Fx(t)$ with $F := [C C \in \mathbb{R}^{1 \times 3}]$. Conversely, $x(t)$ is determined by $\tilde{y}(t; x(0))$ if $D$ is large enough.

**Lemma 8:** Let $\beta = 2\|A\|^2 + \|B\|^2/\|C\|^2_2 \kappa(a)$. For any $T \geq 0$, if $D > 2(n-1) + 2\beta$, then $F$ is column full rank and $x(t) = F^+ \tilde{y}(t; x(0))$ with $F^+ = (F^\top F)^{-1} F^\top$.

**Proof:** Let $\beta = \max_i \|\lambda_i(A_{K_t} - \lambda_i(A_{K})\|$. Theorem 2 in [15] and the assumption that $(A, C)$ is observable imply that $D > 2(n-1) + 2\beta$, $F$ is column full rank. Therefore, $F^+$ is well defined and

$$
(F^\top F)^{-1} F^\top \tilde{y}(t; x(0)) = (F^\top F)^{-1} F^\top Fx(t) = x(t).
$$

Thus, it is sufficient to show $\beta \geq \beta'$. We have

$$
\beta' \leq 2 \max_i \|\lambda_i(A_{K_t})\| \leq 2\|A_{K}\|_2 \leq \beta,
$$

which completes the proof.

As a corollary, we can show $\tilde{f}(K; x(t))$ can be expressed as a quadratic form in terms of $\tilde{y}(t; x(0))$.

**Corollary 1:** For any $x(0)$,

$$
\tilde{f}(K; x(t)) = \tilde{y}(t; x(0))^\top P(K) \tilde{y}(t; x(0)),
$$

where $P(K) = (F^+)^\top X F^+$.

**Proof:** Let $v_j = x_j(0) - x_j(s) x_j(s)^\top$ and $w_j = \tilde{y}(0; x_j(0)) - \tilde{y}(s; x_j(0))^\top$. We define $V$ by the linear space generated by $\{v_j\}$, and $W$ by the linear space generated by $\{w_j\}$. Since $V \in S^2$ and $\dim V = \dim S^2 = \frac{n(n+1)}{2}$, we have $V \in S^n$. Thus, the set $\{v_j\}$ is a basis of $S^n$. Let $v = x(t)x(t)^\top$ for $x(t) \in \mathbb{R}^n$. Then, there exists the sequence $(\alpha_i)$ such that $v = \sum_{i=1}^{\frac{n(n+1)}{2}} \alpha_i v_i$. Define the linear map $F: V \rightarrow W$ by $F(v') = Fv'$, where $v' \in V$. Note that $F(v_j) = w_j$. From (24), we have $\tilde{P}(K), F(v_j)) = (P(K), F(v_j))$, and thus $F(K; F(v)) = (P(K), F(v))$. Then, (32) yields

$$
\tilde{P}(K, F(v)) = \tilde{y}(t; x(0))^\top P(K) \tilde{y}(t; x(0)) = \tilde{f}(K; x(t)).
$$

Therefore,

$$
\tilde{y}(t; x_0)^\top \tilde{P}(K) \tilde{y}(t; x_0) = \langle \tilde{P}(K), F(v) \rangle = \tilde{f}(K; x(t)),
$$

which completes the proof.

Since $\tilde{f}(K; x(0)) = \tilde{f}(K; x(t))$, Corollary 1 and Lemma 9 ensure that Theorem 5 holds.

**References**

[1] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi, “Global convergence of policy gradient methods for the linear quadratic regulator,” in Proc. Int. Conf. Mach. Learn., 2018, pp. 1467–1476.

[2] H. Mohammadi, A. Zare, M. Soltanolkotabi, and M. R. Jovanovic, “Convergence and sample complexity of gradient methods for the model-free linear quadratic regulator problem,” IEEE Trans. Autom. Control, vol. 67, no. 5, pp. 2435–2450, May 2022.
[3] Y. Park, R. Rossi, Z. Wen, G. Wu, and H. Zhao, “Structured policy iteration for linear quadratic regulator,” in Proc. Int. Conf. Mach. Learn., 2020, pp. 7521–7531.

[4] B. Hambly, R. Xu, and H. Yang, “Policy gradient methods for the noisy linear quadratic regulator over a finite horizon,” SIAM J. Control Optim., vol. 59, no. 5, pp. 3359–3391, 2021.

[5] J. Bu, A. Mesbahi, and M. Mesbahi, “LQR via first order flows,” in Proc. Amer. Control Conf., Denver, CO, USA, 2020, pp. 4683–4688.

[6] S. Talebi and M. Mesbahi, “Riemannian constrained policy optimization via geometric stability certificates,” in Proc. IEEE 61st Conf. Decis. Control, Cancun, Mexico, 2022, pp. 1472–1478.

[7] P. Makila and H. Toivonen, “Computational methods for parametric LQ problems—A survey,” IEEE Trans. Autom. Control, vol. 32, no. 8, pp. 658–671, Aug. 1987.

[8] I. Fatkhullin and B. Polyak, “Optimizing static linear feedback: Gradient method,” SIAM J. Control Optim., vol. 59, no. 5, pp. 3887–3911, 2021.

[9] L. M. Zhu, H. Modares, G. O. Peen, F. L. Lewis, and B. Yue, “Adaptive suboptimal output-feedback control for linear systems using integral reinforcement learning,” IEEE Trans. Control Syst. Technol., vol. 23, no. 1, pp. 264–273, Jan. 2015.

[10] B. Jacob and H. Zwart, Linear Port-Hamiltonian Systems on Infinite-dimensional Spaces. Berlin, Germany: Springer, 2012.

[11] M. R. Jovanović and N. K. Dhingra, “Controller architectures: Tradeoffs between performance and structure,” Eur. J. Control, vol. 30, pp. 76–91, 2016.

[12] R. Bellman, “Notes on matrix theory—x a problem in control,” Quart. Appl. Math., vol. 14, no. 4, pp. 417–419, Jan. 1957.

[13] H. Feng and J. Lavaei, “On the exponential number of connected components for the feasible set of optimal decentralized control problems,” in Proc. Amer. Control Conf., 2019, pp. 1430–1437.

[14] S. Prajna, A. V. D. Schaft, and G. Meinsma, “An LMI approach to stabilization of linear port-controlled Hamiltonian systems,” Syst. Control Lett., vol. 45, no. 5, pp. 371–385, 2002.

[15] H. Modares, F. L. Lewis, and Z.-P. Jiang, “Optimal output-feedback control of unknown continuous-time linear systems using off-policy reinforcement learning,” IEEE Trans. Cybern., vol. 46, no. 11, pp. 2401–2410, Nov. 2016.

[16] S. A. A. Rizvi and Z. Lin, “Output feedback reinforcement learning control for the continuous-time linear quadratic regulator problem,” in Proc. Amer. Control Conf., 2018, pp. 3417–3422.

[17] A. Beck, First-Order Methods in Optimization. Philadelphia, PA, USA: SIAM, 2017.

[18] S. Ghadimi, G. Lan, and H. Zhang, “Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization,” Math. Program., vol. 155, no. 1–2, pp. 267–305, Jan. 2016.

[19] I. Grondman, L. Busoniu, G. A. D. Lopes, and R. Babuska, “A survey of actor-critic reinforcement learning: Standard and natural policy gradients,” IEEE Trans. Syst., Man Cybern. Part C, Appl. Rev., vol. 42, no. 6, pp. 1291–1307, Nov. 2012.

[20] T. Degris, P. M. Pilarski, and R. S. Sutton, “Model-free reinforcement learning with continuous action in practice,” in Proc. Amer. Control Conf., 2012, pp. 2177–2182.

[21] V. Mnih et al., “Asynchronous methods for deep reinforcement learning,” in Proc. Int. Conf. Mach. Learn., 2016, pp. 1928–1937.

[22] S. M. Kakade, “A natural policy gradient,” in Proc. Int. Conf. Adv. Neural Inf. Process. Syst., vol. 14, 2001, pp. 1–8.

[23] J. A. Tropp, “User-friendly tail bounds for sums of random matrices,” Foundations Comput. Math., vol. 12, no. 4, pp. 389–434, 2012.