A Low Complexity Approach to Model-Free Stochastic Inverse Linear Quadratic Control

SHANELLE G. CLARKE\(^1\), SOOYUNG BYEON\(^2\), AND INSEOK HWANG\(^3\)

\(^1\)School of Aeronautics and Astronautics, Purdue University, West Lafayette, IN 47907 USA (e-mail: clarke24@purdue.edu)
\(^2\)School of Aeronautics and Astronautics, Purdue University, West Lafayette, IN 47907 USA (e-mail: sbyeon@purdue.edu)
\(^3\)School of Aeronautics and Astronautics, Purdue University, West Lafayette, IN 47907 USA (e-mail: ihwang@purdue.edu)

ABSTRACT In this paper, we present a Model-Free Stochastic Inverse Optimal Control (IOC) algorithm for the discrete-time infinite-horizon stochastic linear quadratic regulator (LQR). Our proposed algorithm exploits the richness of the available system trajectories to recover the control gain \(K\) and cost function parameters \((Q, R)\) in a low (space, sample, and computational) complexity manner. By leveraging insights on the stochastic LQR, we guarantee well-posedness of the Model-Free Stochastic IOC LQR via satisfaction of the Certainty Equivalence optimality conditions. The exact solution of the control gain \(K\) is recovered via a deterministic, low complexity Least Squares approach. Using \(K\), we solve a completely model-free non-iterative SemiDefinite Programming (SDP) problem to obtain a unique (up to a scalar ambiguity) \((Q, R)\), in which optimality and feasibility are jointly ensured. Via derivation of the sample complexity bounds, we show that the non-asymptotic performance of the Model-Free Stochastic IOC LQR can be characterized by the signal-to-noise (SNR) ratio of the finite set of system state and input signals. We present a model-based version of the algorithm for the special case where \((A, B)\) is available, and we, further, provide the extension to the Stochastic Model-Free IOC linear quadratic tracking (LQT) case.

INDEX TERMS Inverse Optimal Control, Knowledge Acquisition, Linear Systems, Optimization, SemiDefinite Programming (SDP), Statistical Learning

I. INTRODUCTION

THE inverse optimal control (IOC) problem was first formulated in 1964 by Kalman [1] for the deterministic single-input linear quadratic regulator (LQR). The extension to the multi-input case was later formulated in [2]. Since then, research interests in studying the IOC problem have been strongly motivated by the variety of applications that require an understanding of “optimal behaviors” [3], [4]. By reverse-engineering the control policy and cost function from the system trajectories, we can gain a greater understanding on how individuals make decisions to better inform system (and individual) behavioral models, which allows us to make better predictions.

Historically, many IOC LQR works have focused on recovering the cost function parameters \((Q, R)\) for a known stabilizing gain \(K\) [2], [5], [6]. More recently, however, the assumption on the knowledge of the control gain \(K\) has been relaxed, and emphasis has been placed on the well-posedness of the IOC LQR problem, i.e., the feasibility of the IOC LQR, and the existence and uniqueness of the feasible solution [7]. Not only does ill-posedness make it difficult to numerically solve the IOC LQR, but also even if there exists a solution, non-uniqueness of the solution does not allow for useful inferences on the decision-making process [8]. Many papers have therefore established feasibility [8], [9] and sufficient conditions on uniqueness [10]–[12] via the use of the Karush-Kuhn-Tucker (KKT) optimality conditions. Feasibility and uniqueness conditions are well established for the deterministic IOC LQR, which is more well-studied than the stochastic case. This is unsurprising as it turns out that it is difficult to demonstrate well-posedness if the assumption that the data is noiseless no longer holds. However, the solution of the deterministic IOC LQR was proven to be statistically consistent to the stochastic case for zero-mean noise [9]. Still, the proofs on the well-posedness of the IOC problem in [9] are only shown for the ideal noiseless case. Proofs on the sufficient uniqueness conditions of the solution \((Q, R)\) for the stochastic IOC LQR, using zero-mean noisy data, are...
presented in [10] for the infinite-horizon case, and in [12] for the finite-horizon case).

Related works, in inverse reinforcement learning (IRL), typically solve the stochastic IOC LQR using a direct approach [13]–[16], while traditional IOC LQR methods use model-based approaches. Approaches in which the nominal model (in our case, the system matrices \((A, B)\)) is assumed to be completely unknown are called direct / model-free approaches. Correspondingly, indirect or model-based approaches either estimate or assume the nominal model to be known a-priori. Model-free approaches have the following distinct strengths: 1) model-free approaches are better able to capture parametric uncertainties and encapsulate the full system behavior that may remain unmodeled owing to data-fitting methods [17]; 2) it has been posited that directly learning a task is often easier than fitting a general purpose system model [18]; and 3) model-based methods tend to achieve more conservative asymptotic behavior than model-free techniques owing to compounding model errors [17].

The lack of model knowledge, in direct approaches, further exacerbates the challenge of establishing well-posedness of the stochastic IOC LQR. Moreover, since model-free approaches utilize a finite number of data points, non-asymptotic properties must be derived in order to show performance guarantees [19]. In general, stochastic IRL works, such as [13]–[15], [20], tend to suffer from ill-posedness and nonconvexity issues [7]. In addition, few IRL works give comments on the sample and computational complexities of their algorithms that require large quantities of data [21], [22]. We cite the works in some of the most recent IRL papers to highlight this gap in the literature. Sample complexity bounds for general IRL methods were presented in both [22] and [23] without comments on well-posedness; and the sufficient conditions on the solution uniqueness, without a feasibility analysis, were presented in [16]. No derivations or comments on the non-asymptotic properties of the proposed method in [16] were provided. Additionally, the method in [16] requires either a-priori knowledge or a guess on \(R\) to estimate \(Q\) (which may then be inaccurate).

In this paper, we propose a Model-Free Stochastic IOC LQR algorithm for the linear time-invariant (LTI) system subject to additive zero-mean Gaussian noise. The contributions of this paper are three-fold:

1) A Well-Posed Model-Free Stochastic IOC LQR algorithm: the proposed algorithm is a SemiDefinite Programming (SDP) optimization problem whose well-posedness is proven via satisfaction of the Certainty Equivalence KKT optimality conditions which establishes feasibility and uniqueness of a solution up to a scalar ambiguity.

2) Low Complexities: the proposed method derives the finite bounds of the sample and space complexities of the proposed algorithm for the LTI system with a \(n\)-dimensional state space and \(m\)-dimensional control input space. The proposed algorithm is also prescribed a worst-case polynomial-time computational complexity.

3) A variant Model-based Stochastic IOC LQR algorithm: we provide a variant of the Model-Free Stochastic IOC LQR algorithm that incorporates the system parameters, if available, to formulate a stochastic model-based algorithm that is also well-posed.

This paper is organized as follows: Section II presents the problem formulation of the infinite-horizon discrete-time stochastic IOC LQR. Section III presents the Model-Free Stochastic IOC LQR algorithm with proofs on well-posedness and low complexities. In Section IV, we present the Model-based Stochastic IOC LQR algorithm, a variant of the original model-free version. In Section V, we demonstrate the performance of the proposed model-free algorithm via simulation. Here, we also compare the performance of the model-based variant algorithm with other existing methods in [8] and [10]. Finally, conclusions are drawn in Section VI, and the extensions to the model-based and Stochastic IOC LQT case are presented in Appendix A.

II. PROBLEM FORMULATION

Consider the discrete-time linear time-invariant (LTI) system:

\[
x_{k+1} = Ax_k + Bu_k + w_k,
\]

where \(x_k \in \mathbb{R}^n\) and \(u_k \in \mathbb{R}^m\) are the respective state and input vectors. \((A \in \mathbb{R}^{n \times n}, B \in \mathbb{R}^{n \times m})\) is the pair of unknown system matrices. The external mismatched disturbance term \(w_k \in \mathbb{R}^n \sim N(0, \Sigma_w)\) is the zero-mean Gaussian noise with a covariance \(\Sigma_w \succ 0\). In this formulation, all state variables are fully observed, and the initial condition \(x_0 \in \mathbb{R}^n \sim N(0, X_0)\) with mean \(\mathbb{E}(x_0) = 0\) and (possibly unknown) covariance \(X_0 \succeq 0\), without loss of generality. The notations, \(M \succ 0\) and \(M \succeq 0\), denote a positive definite and semi-positive definite matrix, \(M\), respectively.

For the stochastic infinite-horizon linear quadratic regular (LQR) in optimal control theory, a stabilizing optimal control policy \(u_k = Kx_k\), \(\forall k \geq 0\), is used to minimize a cost function,

\[
J_0 \triangleq \lim_{T \to \infty} \frac{1}{T} \mathbb{E} \left[ \sum_{k=0}^{T-1} (x_k^T Q x_k + u_k^T R u_k) \right],
\]

parametrized by weights \(Q \in \mathbb{R}^{n \times n} \succeq 0\) and \(R \in \mathbb{R}^{m \times m} \succ 0\) for the state and input terms, respectively. Conversely, the stochastic inverse optimal control (IOC) LQR problem is defined by the following two subproblems [6]. Given a control policy \(u_k = Kx_k\), \(\forall k \geq 0\), that generates a set \(\Gamma\) of noisy trajectories of (1).

Subproblem P1: Recover the gain \(K \in \mathbb{K}_{A,B}\) such that \(K\) is the stabilizing optimal control gain that is necessary and sufficient to minimize the cost function in (2).

Subproblem P2: Determine the pair of unknown weights \((Q, R)\) such that the cost function (2) is minimized for every trajectory in \(\Gamma\).
We denote $\mathcal{K}_{A,B}$ to be the set of all stabilizing feedback gains such that, for $K \in \mathcal{K}_{A,B} \subset \mathbb{R}^{m \times n}$, the spectral radius $\rho(A + BK) < 1$. To facilitate our discussion, we state the assumptions that are made throughout the rest of this paper.

**Assumption 1.**

1. $(A, B)$ is stabilizable and $(A, Q^2)$ is detectable.
2. a) There is a set of $r \geq n + m$ state and input trajectories $\Gamma \triangleq \{(x_{k_1}, u_{k_1}), \ldots, (x_{N_1}, u_{N_1})\}$, for $j \in \{1, \ldots, r\}$ available for a fixed control policy $u_k = Kx_k, \forall k > k_j$, with varying starting conditions $v_{k_j} \triangleq [x_{k_j}, u_{k_j}]^T \in \mathbb{R}^{n+m} \sim \mathcal{N}(0, \Sigma_v)$, and lengths $T_j \triangleq (N_j - k_j + 1) \geq 2$.
   b) The set of trajectories $\Gamma$ is sufficiently excited (rich) so that $\left(\sum_{r=1}^{r} \sum_{k=k_j}^{N_j} x_k u_k^T \right) \succ 0$.

$\Sigma_v$ is the covariance of the starting condition $v_{k_j}$ and we impose that $||\Sigma_v||_2 \geq \left[\left[I_n \quad K^T\right] X_0 \left[I_n \quad K\right]\right]_2$, $I_n \in \mathbb{R}^{n \times n}$ is the identity matrix of size $n \times n$. Assumption 1.1) is standard in LQR applications [8], [24] and allows for the existence of a feasible solution to the (stochastic) LQR problem. Under Assumption 1.2a), unlike the methods in [9], [12], our proposed method allows for the use of multiple partial (or incomplete) trials of state and input information. As such, the trajectories in $\Gamma$ can have varying lengths $T_j$ and starting conditions $v_{k_j}$. The benefit of which is that there is no need for uniform trajectory lengths or fixed starting conditions, and extensive information preprocessing is rendered unnecessary.

Assumption 1.2a) also facilitates and allows Assumption 1.2b) to hold. Assumption 1.2b) requires that the collected trajectories must be rich enough such that the system information varies sufficiently to encapsulate the entire behavior of the data-generating dynamical system in (1). This assumption is mild and common in data-driven/model-free approaches [25], [26]; and is related to the idea of Persistence of Excitation [26], [27]. To ensure that this condition on richness is met, the system trajectories are excited via two means: (i) varying the starting condition $v_{k_j}$ (as per Assumption 1.2a), and (ii) injecting the zero-mean white noise process $w_k$. Both excitive means (i) and (ii) encourage exploration of the state space. We refer readers to [26] and [27] for a treatise on Persistence of Excitation in data-driven control methods.

**Remark 1.** We must remark that our proposed approach has a lower space complexity over the methods in [26] and [16], in which the dimensions of the relevant (Hankel) data matrices increase as the number $r$ and the length $T$ of trajectories increase. In our approach, the associated data matrix, in Assumption 1.2b) above, maintains a fixed dimension $(n + m) \times (n + m)$, irrespective of $r$ and $T$.

III. THE INFINITE-HORIZON MODEL-FREE STOCHASTIC IOC LQR

As previously discussed in Subproblems P1 and P2 respectively, we are interested in 1) recovering the stabilizing optimal control gain $K$, and 2) computing $(Q, R)$ in a model-free manner since we only have access to a set of rich trajectories $\Gamma$. In this section, we present the proposed Model-Free Stochastic IOC LQR method and its corresponding algorithm (Algorithm 1) that is comprised of two main parts. The first part, in Section III-A, extracts an exact solution of $K$ from the system trajectories, $\Gamma$. We prove that we can efficiently and exactly recover $K$ from as little as $n + m$ rich trajectories, each of, at least, length $2$. The second part, in Section III-B, presents the Model-Free Stochastic IOC LQR SDP formulation with a rigorous proof on its well-posedness and sample complexity bounds. The Model-Free Stochastic IOC LQR algorithm (Algorithm 1) generates a unique certainty-equivalence solution $(Q, R)$ that is cheap in terms of its space, sample and computational complexities.

A. EXTRACTION OF THE CONTROL GAIN $K$

In order to extract an exact solution for the control gain $K$, consider the data matrix $X_r^1 > 0$ defined as:

\[
X_r^1 \triangleq \frac{1}{r} \sum_{j=1}^{r} \left( \frac{1}{T_j - 1} \sum_{k=k_j+1}^{N_j} \left[ x_k u_k^T \right] \right)
\]

Going forward, we define the partition of any matrix $M \in \mathbb{R}^{(n+m) \times (n+m)}$ as $M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}$ where $M_{11} \in \mathbb{R}^{n \times n}$, $M_{12} \in \mathbb{R}^{n \times m}$, $M_{21} \in \mathbb{R}^{m \times n}$, and $M_{22} \in \mathbb{R}^{m \times m}$. $X_r^1$ can be viewed as a measure of the sample covariance of the system data from a zero-mean equilibrium point. In this work, we use $X_r^0$ as the marginal sample covariance as it incorporates the starting conditions $v_{k_j}$. Remember that varying the starting conditions encourages state-space exploration and as a result, $X_r^0$ gives the best measure of the sample covariance. Since the trajectories, $\Gamma$, are propagated using a control law $u_k = Kx_k, \forall k > k_j$, from some starting condition $v_{k_j}$, the one-step forward sample covariance, $X_r^1$, is directly parametrized in terms of $K$, and can be used to extract $K$ exactly as $K = (X_r^1)^T_{12}(X_r^1)^{-1}_{11}$, as shown in Lemma 1.

**Lemma 1.** For $X_r^1 > 0$ and $(X_r^1)^{11} > 0$, $K = K_r^1 = (X_r^1)^T_{12}(X_r^1)^{-1}_{11}$ and $K_r^1$ is unique.

**Proof.** We examine the partitions of $X_r^1$:

\[
X_r^1 = \frac{1}{r} \sum_{j=1}^{r} \left( \frac{1}{T_j - 1} \sum_{k=k_j+1}^{N_j} \left[ I_n \quad K \right] x_k x_k^T \left[ I_n \quad K \right]^T \right) = \left[ I \quad K \right]^T (X_r^1)^{11} \left[ I \quad K \right]^T
\]

We see that $(X_r^1)^T_{12} = K(X_r^1)^{11}$. Hence, $K = K_r^1 = (X_r^1)^T_{12}(X_r^1)^{-1}_{11}$ and $K_r^1$ is unique for $(X_r^1)^{11} > 0$. \(\square\)
The computation of $K$ from (4) may seem trivial. As the reader may suspect, it is a linear least-squares computation. However, the contribution here is not an artifact of the computational method used but rather from the fact that, by exploiting the richness of the data, the computation of $K$ is a linear matrix equality system that is consistent and has exactly one solution. Therefore, we can compute $K$ in a deterministic low (computation) complexity manner that leaves an exactly zero residual, $\epsilon_K \triangleq ||K-K_0||_F = 0$. $||M||_F$ denotes the Frobenius norm of the matrix $M$. The matrix space of $X_1^i$ is made of linear combinations of the matrices $V_{k_j}$ and $H_{1,j}$ for $j \in \{1, \ldots, r\}$. Thus, for a non-zero matrix $A_K$, we need at least $n+m$ independent $v_{k_j}$ or $n+m$ independent $w_{k_j}$ so that $\text{rank}(X_1^i) = n+m$ and $X_1^i \succ 0$. Additionally, $X_1^i \succ 0$ if and only if all its leading principal minors are positive. Hence, $(X_1^i)_{11} \succ 0$.

This result makes intuitive sense. Where the number of predictors (resp. dimension of $X_1^i$) is of size $n+m$, we need at least $n+m$ observations (resp. number of trajectories). This is a classic result from the literature on linear regression system identification techniques [29]. We can also talk about what is the minimum length of a single trajectory such that $X_1^i \succ 0$ via Corollary 1.

**Corollary 1.** For a set, $\Gamma$, containing a single trajectory, i.e. $r = 1$, we need a trajectory of length $T_r \geq n+m+1$, excited by a sequence of independent $n+m$ noise vectors $w_{k_j}$, so that $X_1^i \succ 0$ and $(X_1^i)_{11} \succ 0$.

**Remark 2.** We see that the condition $X_1^i \succ 0$ can be met using a single trajectory, but this requires a trajectory of length $T_r \geq n+m+1$ with a sequence of $n+m$ independent noise vectors that excites all the modes of the system. However, it is arguably easier to vary the system starting conditions than to excite all the modes of a unknown dynamical system using noise.

## B. STOCHASTIC MODEL-FREE IIOC LQR SDP OPTIMIZATION

Given the exact computation of the control gain $K$, we present the Stochastic Model-Free IIOC LQR SDP Optimization problem, $P_{\text{inv}}$, in (7), and claim that $P_{\text{inv}}$ is well-posed and obtains unique solutions, $(\hat{Q}_{\text{inv}}, \hat{R}_{\text{inv}})$, within a scalar ambiguity of $(Q, R)$. We prove these claims via Theorems 1 and 2. Consider the Stochastic Model-Free IIOC LQR SDP Optimization problem:

$$P_{\text{inv}} : \begin{cases} \sup_{P, \hat{Q}, \hat{R}} \frac{1}{T_r} \sum_{j=1}^{T_r} (\frac{1}{T_r} \sum_{k=0}^{T_r} x_{k+1}^T \Sigma_r^{-1} x_{k+1} + \frac{1}{T_r} \sum_{k=0}^{T_r} u_{k+1}^T \Sigma_r^{-1} u_{k+1} ) \geq 0, \\ P_{12} + (K_0^i)^T P_{22} = 0, \\ Z^T P Z + X_0^i (\hat{\Lambda} - P) X_0^i = 0, \\ P_{22} > 0, Q > 0, \hat{R} > 0, \end{cases}$$

where $P \in \mathbb{R}^{(n+m) \times (n+m)}$ is the Lyapunov SDP variable, $\hat{\Lambda} \triangleq \frac{1}{T_r} \sum_{j=1}^{T_r} \sum_{k=j}^{T_r} x_{k+1}^T \Sigma_r^{-1} x_{k+1} + \frac{1}{T_r} \sum_{k=j}^{T_r} u_{k+1}^T \Sigma_r^{-1} u_{k+1}$, and $\hat{\Lambda} \triangleq \text{blkdiag}(\hat{Q}, \hat{R})$ is the estimate of the true $\Lambda \triangleq \text{blkdiag}(Q, R)$. $\text{blkdiag}(C, D)$ denotes a block diagonal matrix with matrices $C$ and $D$. We investigate the well-posedness of $P_{\text{inv}}$ via a two-pronged approach. First, we prove that $P_{\text{inv}}$ is feasible by establishing that there exists a non-empty set of solutions that satisfy the constraints in (7), and that, despite the lack of upperbound on $P$ in (7), $J_{\text{inv}}$ is bounded. We leverage insights on the feasibility of the stochastic LQR to inform that of the stochastic IOC LQR; and then, solve the feasible stochastic IIOC LQR using a
Certainty Equivalence approach. The Certainty Equivalence approach involves adopting the estimate of a parameter as if it were the true parameter \cite{30}. We refer readers to \cite{30, 31} for further exposition.

Consider the stochastic LQR problem with the LTI dynamics (1) and the cost function (2). Let \( \Omega_0 \) be the set of solutions of the stochastic LQR, given a pair of parameters \((Q, R)\) belonging to a set \( \Omega_{Q, R} \). In the stochastic LQR setting under Assumption 1.1, it is commonly known that there exists an optimal solution, \( J_0 = \text{Trace}(X_0 + \Sigma w)S^* \), given a linear control policy \( u_k = K^* x_k, \forall k \geq 0 \), where the optimal stabilizing gain \( K^* \in K_{A,B} = -(R + B^T S^* B)^{-1} B^T S^* A \). \( S^* \succeq 0 \) is the unique solution of the Discrete Algebraic Riccati Equation (DARE):

\[
S^* = Q + A^T S^* A - A^T S^* B(R + B^T S^* B)^{-1} B^T S^* A.
\]

Therefore, \((K^*, S^*) \in \Omega_0 \) and \( \Omega_0 \) is non-empty. The stochastic LQR problem is equivalent to:

\[
P_1 : \inf_{X,K} J_{X,K} \triangleq \text{Trace}(AX) \text{ subject to: } X = AKX A_K^T + \Sigma_K, \ X > 0,
\]

with the parametrization:

\[
X_T = \frac{1}{T} \mathbb{E} \left( \sum_{k=0}^{T-1} \begin{bmatrix} x_k \\ u_k \end{bmatrix} \begin{bmatrix} x_k \\ u_k \end{bmatrix}^T \right), \quad \Sigma_{K,T} = \frac{1}{T} \mathbb{E} \left( \sum_{k=0}^{T-1} I_n K \begin{bmatrix} I_n \\ K \end{bmatrix}^T + \begin{bmatrix} x_0 \\ u_0 \end{bmatrix} \begin{bmatrix} x_0 \\ u_0 \end{bmatrix}^T \right),
\]

\( X \triangleq \lim_{T \to \infty} X_T \), and \( \Sigma_K \triangleq \lim_{T \to \infty} \Sigma_{K,T} \) are the time-averaged steady state covariances of the augmented state and input vector, and the excitive inputs, respectively. We use Lemma 2 to prove that the primal problem \( P_1 \) is equivalent to the LQR problem by \((J_{X,K}^*, K_{X,K}^*) \equiv (J_0^*, K^*)\). Let the tuple \((J_{X,K}^*, K_{X,K}^*, X_{K}^*)\) be the optimal solution of the primal \( P_1 \).

**Lemma 2.** \((J_{X,K}^*, K_{X,K}^*) \equiv (J_0^*, K^*)\).

**Proof.** For any stabilizing \( K \in K_{A,B}, \) the cost function \( J_0 \) is equivalently written as

\[
J_0 = \lim_{T \to \infty} \frac{1}{T} \mathbb{E} \left( \sum_{k=0}^{T-1} \begin{bmatrix} x_k \\ u_k \end{bmatrix} \right)^T \Lambda \begin{bmatrix} x_k \\ u_k \end{bmatrix} = \text{Trace}(AX).
\]

Consider the sum of state and input trajectories propagated via the dynamics in (1):

\[
\sum_{k=0}^{T-1} \begin{bmatrix} x_{k+1} \\ u_{k+1} \end{bmatrix} = \sum_{k=0}^{T-1} \begin{bmatrix} A_K \begin{bmatrix} x_k \\ u_k \end{bmatrix} + \begin{bmatrix} I_n \\ K \end{bmatrix} w_k \end{bmatrix}.
\]

Then, given all covariance terms \( \mathbb{E}(x_k u_k^T) = \mathbb{E}(u_k x_k^T) = 0 \),

\[
\mathbb{E} \left( \sum_{k=0}^{T-1} \begin{bmatrix} x_k \\ u_k \end{bmatrix} \begin{bmatrix} x_k \\ u_k \end{bmatrix}^T \right) = A_K \mathbb{E} \left( \sum_{k=0}^{T-1} \begin{bmatrix} x_k \\ u_k \end{bmatrix} \begin{bmatrix} x_k \\ u_k \end{bmatrix}^T \right) A_K^T
\]

\[
+ \mathbb{E} \left( \sum_{k=0}^{T-1} \begin{bmatrix} I_n \\ K \end{bmatrix} w_k w_k^T \begin{bmatrix} I_n \\ K \end{bmatrix} + \begin{bmatrix} I_n \\ K \end{bmatrix} x_0 x_0^T \begin{bmatrix} I_n \\ K \end{bmatrix} \right).
\]

We take the \( \lim_{T \to \infty} \) of the time average of (13) to find the time-averaged steady state formulation, which is the matrix equality constraint as in (9). The stochastic LQR problem also requires that the gain \( K \) be stabilizing. By Proposition [25, Lemma 1, pg. 3757], \( \rho(A + BK) = \rho(AK) \). Thus, \( K \in K_{A,B} \) is equivalent to \( \rho(AK) < 1 \). By the Lyapunov Stability Lemma [32, Thm. 3.18, pg. 85], given \( \Sigma_K > 0 \), \( \rho(AK) < 1 \) if and only if \( X > 0 \); and also \( X \) is unique. Thus, \( X_{K}^* \) is unique, \( K_{X,K}^* = K^* \) is unique, and \( J_{X,K}^* = J_0^* \).

From \( P_1 \), we can analyze the feasibility of the stochastic LQR (and thereby, the feasibility of the stochastic IOC LQR) from consideration of its Karush-Kuhn-Tucker (KKT) conditions. The first-order necessary KKT conditions of \( P_1 \) are:

- **Primal Feasibility:** \( AXKA_K^T - X + \Sigma_K = 0, X > 0 \),
- **Stationarity:** \( \frac{\partial L}{\partial X} = \lambda - P + A_K^T PA_K = 0 \), \( \frac{\partial L}{\partial K} = 2A_b X A_b^T (P_I + K^T P_22) = 0 \),
- **Complementary Slackness:** \( -Y = X \), \( Y \geq 0 \),
- **Dual Feasibility:** \( P \succeq 0 \),

and the second-order sufficient KKT conditions are:

\[
\frac{\partial^2 L}{\partial X^2} = 0, \quad \frac{\partial^2 L}{\partial K^2} = 2A_b X A_b^T P_22 > 0,
\]

where \( A_b = [\begin{bmatrix} A & B \end{bmatrix}] \). \((P, Y)\) is the pair of Lagrangian dual variables of the respective equality and inequality conditions in (9). Given that there exists a non-empty \( \Omega_0 \) for a fixed \( \Omega_{Q,R} \) that satisfies the KKT conditions, the KKT conditions can be used to find the feasible \( \Omega_{Q,R} \) for a fixed non-empty \( \Omega_0 \). We use Theorem 1 to prove that the Stochastic Model-free IOC LQR problem, \( P_{\text{inv}} \) is feasible. We set \( X_0^* \) as the (so-called) minimum-stress estimate of \( X \) by the Certainty Equivalence Principle.

**Lemma 3.** For a non-empty set \( \Omega_0 \) containing \((K, X)\), the set of solutions \( \Omega_{Q,R} \) to \( P_{\text{inv}} \) is non-empty and contains a feasible \((Q_{\text{inv}}^*, R_{\text{inv}}^*, P_{\text{inv}}^*)\).

**Proof.** By construction, \( X_0^* > 0 \) satisfies the conditions on primal feasibility. \( Z^T PZ + X_0^*(\Lambda - P)X_0^* = 0 \) is the Certainty Equivalence formulation of \( X(A - P + A_K^T PA_K)X = 0 \), which satisfies the first stationarity condition in (14). The constraint \( P_{12} + K^T P_{22} = 0 \) satisfies the second condition on stationarity. For satisfaction of optimality and dual feasibility, we respectively include the sufficient condition \( P_{22} > 0 \) and \( P \succeq 0 \). The condition on complementarity is satisfied for \( Y = 0 \). Therefore, \( P_{\text{inv}} \) satisfies the necessary and sufficient

\[ \text{null.} \]
KKT conditions. Given that \((K, X) \in \Omega_0\) and the stochastic LQR necessarily satisfies (14), \(\Omega_{\hat{Q}, \hat{R}}\) is non-empty.

Now, we prove the cost function \(J_{P_m}\) is bounded.

**Lemma 4.** \(J_{P_{inv}}\) is bounded.

**Proof.** Any solution \(P_{inv}\geq 0\) implies that \(J_{P_{inv}} \geq 0\). For any \(K\) such that \(\rho(A_K) < 1\), the first stationarity condition in (14) ensures that

\[
P = \sum_{i=0}^{\infty} (A_K^T)^i \hat{\Lambda} A_K^i \leq \sum_{i=0}^{\infty} (A_K^T)^i \hat{\Lambda}^j \hat{\Lambda}^i A_K^i
\]

\[
\leq \sum_{i=0}^{\infty} ||\hat{\Lambda}^j|| ||A_K^i||^2 I_{m+n}
\]

\[
\leq \sum_{i=0}^{\infty} ||\hat{\Lambda}^j|| ||A_K^i||^2 I_{m+n}. \tag{16}
\]

Since \(||A_K|| \leq ||A_K||_2\), the upperbound on \(P\) becomes

\[
P \leq ||\hat{\Lambda}^j|| \sum_{i=0}^{\infty} (||A_K||^2)_i I_{n+m}
\]

\[
\leq ||\hat{\Lambda}^j|| \sum_{i=0}^{\infty} (\rho(A_K)^2)_i I_{n+m}
\]

\[
= \frac{||\hat{\Lambda}^j|| ||A_K^i||^2 I_{n+m}}{1 - \rho(A_K)^2} = P_{max}.
\]

Given that the **Certainty Equivalence** solution \(P_{inv}\) satisfies the first stationarity constraint, \(P_{inv}\) is upperbound. \(\square\)

Now, we have all the tools to discuss the feasibility of \(P_{inv}\) using Theorem 1.

**Theorem 1.** \(P_{inv}\) is feasible.

**Proof.** From Lemmas 3 and 4, \(P_{inv}\) obtains a non-empty set of feasible solutions \(\Omega_{\hat{Q}, \hat{R}}\) and \(J_{P_{inv}}\) is bounded between 0 and \(P_{max}\). \(\square\)

**Remark 3.** \(P_{inv}\) is a feasible SDP optimization problem that finds an \(\epsilon\)-approximation solution, using the interior-point method, with worst-case computational complexity linear in \(\log(1/\epsilon)\) and polynomial in \(n + m\) \([33]\).

The second proog in the investigation of well-posedness of \(P_{inv}\) is to prove that the optimal solution tuple \((\hat{Q}_{P_{inv}}, \hat{R}_{P_{inv}})\) is unique. While \((\hat{Q}_{P_{inv}}, \hat{R}_{P_{inv}})\) satisfies the KKT conditions, it is not necessarily true that \((\hat{Q}_{P_{inv}}, \hat{R}_{P_{inv}}) = (Q, R)\). This is because there is a manifold of possible solutions that satisfy the KKT conditions \([8]\). However, by solving (7), we can recover a unique, optimal solution \((\hat{Q}_{P_{inv}}, \hat{R}_{P_{inv}})\) up to a scalar ambiguity, i.e., \((\hat{Q}_{P_{inv}}, \hat{R}_{P_{inv}}) = (\gamma^2 Q, \gamma^2 R)\), where \(\gamma \in \mathbb{R}^1\) is a scalar constant. Theorem 2 proves this result.

**Theorem 2.** \((\hat{Q}_{P_{inv}}, \hat{R}_{P_{inv}}) = (\gamma^2 Q, \gamma^2 R)\) for \(\gamma \in \mathbb{R}^1\).

**Proof.** \(\hat{A} \succeq 0\) is satisfied in (7), \(\rho(A_K) < 1\) and \(Z^T P Z + X_0^T (\hat{\Lambda} - P) X_0 = 0\) is the **Certainty Equivalence** form of \(X(A_K^T P A_K + \hat{\Lambda} - P) X = 0\). \(P \succeq 0\) is unique for \(X \succ 0\) and \(\hat{\Lambda} \succeq 0\), by the Lyapunov Stability Lemma in [25, Lemma 2.1]. The Lyapunov equation is invariant under the similarity transform: \(\hat{A}_K = T A_K T^{-1}\) with solutions \(\hat{P} = T^{-T} P T^{-1}\) and \(\hat{\Lambda} = T^{-T} \Lambda T^{-1}\). This means that all other possible solutions are such that \(X(A_K^T P A_K - \hat{P} + \hat{\Lambda}) X = XT^{-T}(A_K P A_K - P + \hat{\Lambda}) T^{-1} X = 0\). Thus, \(T^{-1} X\) and \(T^{-T} X\) must be commuting since the columns of the transformation matrix \(T^{-1}\) do not provide the eigenbases of the invariant space. \(T^{-1} X\) and \(T^{-T} X\) commute if and only if \(T^{-1}\) is a scaled identity matrix, i.e., \(T^{-1} = \gamma I_{n+m}\) for \(\gamma \in \mathbb{R}^1\). Hence, \((\hat{Q}_{P_{inv}}, \hat{R}_{P_{inv}}) = (\gamma^2 Q, \gamma^2 R)\). \(\square\)

From Theorems 1 and 2, \(P_{inv}\) is well-posed and the proposed Stochastic Model-Free IOC LQR algorithm yields a unique solution. Up to this point, we have investigated the well-posedness of the stochastic IOC LQR via the use of the **Certainty Equivalence Principle**. That is, we have obtained the performance weights \((\hat{Q}, \hat{R})\) by treating the marginal sample covariance \(X_0^0\) as the true covariance \(X\). Although it may be efficient to use such an estimate, it is important to establish the performance gap between the **Certainty Equivalence** form and the true form via the derivation of non-asymptotic performance bounds. We investigate the sample complexity of the minimum stress estimate \(X_0^0\) and provide the error bound \(\epsilon_X \triangleq ||X_0^0 - X||_{max}\) as a measure of the performance of \(P_{inv}\). The smaller the error bound, the smaller the performance gap between \(P_{inv}\) and the non-**Certainty Equivalence** (true) form of \(P_{inv}\); and the smaller the error \(||\Lambda - \hat{\Lambda}||_2\). We provide the sample complexity bounds on \(X_0^0\) using Theorem 3.

**Theorem 3.** For the stochastic LTI system (1) under Assumption I, with a marginal sample covariance \(X_0^0\) defined in (3), we have, with a probability no smaller than \(1 - p\),

\[
\epsilon_X \leq \frac{1}{r} \sum_{j=1}^{r} 16 ||X||_2 \max_{ii} (X_{ii}) \min_{ii} (X_{ii}) (1 - ||A_K||_2)^2,
\]

where

\[
p \triangleq \sum_{j=1}^{r} \left[ 3(n + m)^2 \exp \left( -\frac{T_j + 1}{2} \right) \right. + \left. 3(n + m)^2 \exp \left( -\frac{3(T_j + 1)^2}{2} + (T_j + 1)^2 \right) \right].
\]

The proof of Theorem 3 is given in Appendix B. Theorem 3 offers some interesting insights that we can use to further develop our intuition on how the sample complexity affects the algorithm’s performance. Let us consider the following extreme cases.

- **Case 1:** \(r\) large and all \(T_j\) large such that \(r \approx \infty\) and \(T_j \approx \infty\) for all \(j \in \{1, \ldots, r\}\). Theorem 3 states that with probability 1, the error \(\epsilon_X = 0\). In other words, for an infinite set of trajectories of infinite length, the marginal sample covariance \(X_0^0\) is the true covariance \(X\), and the **Certainty Equivalence** formulation \(P_{inv}\) is the true formulation.
Given online settings, to improve the performance of model-free LQR, the results in our paper, for the model-free stochastic case, there are several associated challenges. Firstly, there is the issue of well-posedness. It is easy to see here that $P_{\text{inv}}$ admits a feasible, unique solution that satisfies the necessary and sufficient KKT conditions in (14) and (15).

Secondly, model-based IOC LQR methods utilize the set of trajectories, $\Gamma$, to estimate the control gain $K$ via some regression technique, and subsequently, solve (8) (or some equivalent optimality equation) to compute $(Q, R)$ given $(A, B)$. Feasibility of the IOC LQR is then particularly sensitive to inexact estimates of $(K, A, B)$ owing to the combinatorial nature of (8). An inaccurate estimate of $K, A$, or $B$ can cause infeasibility so that there exists no solution to (8), as discussed in [8]. Most existing works, therefore, assume that the nominal model is perfectly known, i.e., $(\hat{A}, \hat{B}) = (A, B)$. The proposed method in this paper deals with this issue in two ways:

1) The estimation of $K$ from $K_1$ is exact and admits an error $\epsilon_K = 0$, as discussed in Section III-A. In fact, since there is no longer a need to estimate the sample covariance $X_0^{-1}K_1$ can be exactly estimated from as little as $r = n + m$ trajectories of length 2 or a single trajectory of length $n + m + 1$, as discussed in Proposition 1 and Corollary 1, respectively.

2) If the pair $(A, B)$ is not perfectly known, we can estimate $(\hat{A}, \hat{B})$ using system identification techniques as in [34, Algorithm 1]. (Since these system identification techniques for an LTI system are numerous in the literature, we do not provide such techniques in this paper). We also refer readers to [34, Prop. 1.1] for a sample complexity analysis for estimating the matrices $(\hat{A}, \hat{B})$ using a least-squares regression approach.

We illustrate the comparative performance of the Model-based Stochastic IOC LQR with the methods in [8] and [10] that use stochastic data (in simulation) to solve the Model-based IOC LQR.

IV. A VARIANT ALGORITHM: INFINITE-HORIZON MODEL-BASED STOCHASTIC IOC LQR

Given $(\hat{A}, \hat{B})$, we use Algorithm 2 to solve for an estimate $(\hat{Q}, \hat{R})$. In Algorithm 2, the Certainty Equivalence formulation problem $P_{\text{inv}}$ in (7) is replaced by

$$P_{\text{inv}}^M : \sup_{\hat{Q}, \hat{R}} J_{\text{inv}}^M \triangleq \text{Trace}(P) \text{ subject to:}$$

$$P_{12} + (K_1)^T P_{22} = 0$$

$$\hat{A}_K^T P \hat{A}_K + \hat{A} - P = 0$$

$$P_{22} \succ 0, P \succeq 0, \hat{Q} \succeq 0, \hat{R} \succ 0,$$

where $\hat{A}_K \triangleq [I_n \ (K_1)^T]^T [\hat{A} \ \hat{B}]$. This may seem a somewhat trivial replacement. However, in the Model-based stochastic IOC LQ case, there are several associated challenges. Firstly, there is the issue of well-posedness. It is easy to see here that $P_{\text{inv}}^M$ is well-posed and generates a unique solution $(\hat{Q}_{\text{inv}}^M, \hat{R}_{\text{inv}}^M)$ up to a scalar ambiguity. In a proof similar to Theorem 2, we can guarantee that $P_{\text{inv}}^M$ admits a feasible, unique solution that satisfies the necessary and sufficient KKT conditions in (14) and (15).

Algorithm 2 Model-Based Stochastic IOC LQR

Require: $\Gamma$

Step 1: Construct $X_1^\dagger$ using $\Gamma$

Step 2: Compute $K_1^\dagger = (X_1^\dagger)^T (X_1^\dagger)^{-1}$ from $X_1^\dagger$

Step 3: Solve for $(P, \hat{Q}, \hat{R})$ from (20)

Return $(K_1^\dagger, \hat{Q}, \hat{R})$
$R = 1000I_n$ to illustrate the case where there is a relatively high cost for power consumption to encourage small control inputs. Small control inputs may destabilize a ill-fitted system by generating small inputs for (unstable) modes that are estimated to be stable. The optimal control gain

$$K = \begin{bmatrix} -0.0437 & -0.0125 & -0.0013 \\ -0.0125 & -0.0450 & -0.0125 \\ -0.0013 & -0.0125 & -0.0437 \end{bmatrix}$$  \hspace{1cm} (22)$$
can be obtained by solving the DARE. In this simulation, we solve the set of SDP equations in (7) and (20) using the MATLAB `cvx` SDPT3 solver [36], that has worst-case polynomial time and computational complexity.

**A. NOISY MODEL-FREE CASE**

In the model-free case, the system trajectories are generated with the nominal model ($A, B$) and control gain $K$. In Fig. 1, we demonstrate the Monte-Carlo simulation results of our proposed algorithm in the noisy case. For each episode, the $r$ state and input trajectories were simulated from randomly generated initial points, $v_k \sim \mathcal{N}(0, \text{blkdiag}(10^2I_n, 10^2I_m))$ with $T_j = N$ equal for all trajectories. Figure 1 shows that the accuracy of recovering the cost function increases as the number and length of the trajectories increase. This is to be expected as, with more trajectories of longer lengths, $X^0$ is better able to encapsulate the entire system behavior, and better approximates $X$, as discussed in Section III-B.

We analyze Fig. 1 using the insights gained from Theorem 3. For ease of analysis, we look at the corners of Fig. 1. The lower left-hand corner of Fig. 1 presents the case where $r$ is small and $T_j$ is small, which yields an average error $\|\Lambda - \hat{\Lambda}\|_2 \approx 0.25$. (Remember that for all trajectories, $T_j = N$, in this simulation). Here, the algorithm requires more data points to give a better estimate. In comparison, the upper left-hand corner yields a smaller performance error $\|\Lambda - \hat{\Lambda}\|_2 \approx 0.2$. The upper left-hand corner of Fig. 1 illustrates the case where $r$ is small as compared to $T_j$. Here, the length of the trajectories are longer, which accounts for some performance improvement, but this improvement is tempered by a small SNR. This is because the contribution of the input (noisy) signal, to the covariance, is dominated by that of the process noise. As is expected from our insights on Theorem 3, a small SNR widens the performance gap. The lower right-hand corner outperforms the results from the left half of Fig. 1. Here, the SNR is larger than that of both the lower and upper left-hand corners, as $r$ is large and $T_j$ is small. From Theorem 3, this narrows the performance gap while increasing the probability of success $1 - p$. Finally, the upper right-hand corner illustrates the case where both $r$ and $T_j$ are large. In this corner, the performance results are the best of all the cases.

**B. NOISELESS MODEL-FREE CASE**

For completeness, we present the performance results for the model-free noiseless case using our proposed Model-Free Stochastic IOC LQR algorithm. That is, the excitation of the system stems only from varying the starting points, $v_k \sim \mathcal{N}(0, \text{blkdiag}(10^2I_n, 10^2I_m))$, and $\Sigma_w = 0I_n$. It is obvious here that this is a (ideally) large SNR. For $(T_j = 6, r = 6)$, the normalized error $\|\Lambda - \hat{\Lambda}\|_2 = 3.4 \times 10^{-4}$ and for $(T_j = 300, r = 50)$, $\|\Lambda - \hat{\Lambda}\|_2 = 6.93 \times 10^{-6}$. In the noiseless case, we can even generate good results for the small dataset $(T_j = 2, r = 6)$: $\|\Lambda - \hat{\Lambda}\|_2 = 0.0046$. Overall, the model-free noiseless case shows better comparative accuracy, on the order of $10^{-2}$, to the noisy case since the system data is uncorrupted by noise.

**C. MODEL-BASED CASE**

In this section, we demonstrate the comparative accuracy of our proposed variant model-based algorithm to the LMI-based method in [8] and the inverse KKT approach presented in [10]. Note that the model-based version of our algorithm (Algorithm 2) is employed for this comparison study, since the methods proposed in [8] and [10] are applicable to only the model-based case. Table 1 summarizes the comparative performance results for different scenarios of a varying number of trajectories $r$, lengths of trajectories $T_j$, and noise covariances $\Sigma_w$ for three cases: 1) where $(\hat{A}, \hat{B}) = (A, B)$ is perfectly known, 2) where $(\hat{A}, \hat{B}) = (\hat{A}_1, \hat{B}_1)$ is closed-loop stable for the state-feedback gain $K$, and 3) where $(\hat{A}, \hat{B}) = (\hat{A}_2, \hat{B}_2)$ is closed-loop unstable for the state-feedback gain $K$. Note that, for the relevant cases, $(\hat{A}, \hat{B})$ was estimated using the Least Squares Regression Algorithm in [34, Algorithm 1]. We give these estimated matrices as:

$$
\hat{A}_1 = \begin{bmatrix} 1.0064 & 0.0145 & -0.0029 \\ 0.0035 & 0.9776 & -0.0042 \\ 0.0026 & 0.0227 & 1.0068 \end{bmatrix}, \\
\hat{B}_1 = \begin{bmatrix} 1.0064 & 0.0145 & -0.0029 \\ 0.0035 & 0.9776 & -0.0042 \\ 0.0026 & 0.0227 & 1.0068 \end{bmatrix}, \\
\hat{A}_2 = \begin{bmatrix} 1.0081 & -0.0205 & -0.0383 \\ 0.0143 & 0.9988 & 0.0225 \\ -0.1383 & 0.0090 & 0.9762 \end{bmatrix}, \\
\hat{B}_2 = \begin{bmatrix} 1.0081 & -0.0205 & -0.0383 \\ 0.0143 & 0.9988 & 0.0225 \\ -0.1383 & 0.0090 & 0.9762 \end{bmatrix}.
$$

We provide the eigenvalues of $(\hat{A}_2 + \hat{B}_2 K)$ and $(\hat{A}_2 + \hat{B}_2 K)$ to show that the closed-loop stability for each case. The eigenvalues of $(\hat{A}_1 + \hat{B}_1 K)$ are $\{0.9432, 0.9608, 0.9608\}$ and the eigenvalues of $(\hat{A}_2 + \hat{B}_2 K)$ are $\{0.8791, 1.0329, 0.9511\}$.

Table 1 shows that our algorithm outperforms the others in terms of accuracy across all the scenarios with varying noise, varying number of trajectories of varying lengths, and even varying descriptions of the system models. Compared to [10], our proposed method is well-posed for any given $(\hat{A}, \hat{B})$ and is able to obtain a solution for low sample complexities and data with high noise covariances. The inverse KKT approach in [10] relies on a specific rank condition on the recovery matrix, which is sensitive to the length of the trajectories.
TABLE 1. A comparison of error $||\Lambda - \hat{\Lambda}||_2$ for the model-based approaches proposed in this paper, [8] and [10]

| $\Sigma_w$ | Method | $T_j = 2$ | $T_j = 2$ | $T_j = 2$ | $T_j = 50$ | $T_j = 50$ | $T_j = 50$ | $T_j = 50$ | $T_j = 50$ |
|------------|--------|------------|------------|------------|------------|------------|------------|------------|------------|
|           | Proposed | $1.80 \times 10^{-8}$ | $4.95 \times 10^{-6}$ | $2.34 \times 10^{-7}$ | $0.003$ | $0.003$ | $0.003$ | $0.607$ | $0.607$ |
| $0.001I_n$ | LMI-based [8] | 0.0165 | 0.0149 | 0.0157 | 0.979 | 0.979 | 0.979 | 0.871 | 0.871 |
|           | Inverse KKT [10] | 1.013* | 0.0947 | 0.0263 | 5.173* | 2.103* | 2.023* | 1.016* | 1.124* |
|           | Proposed | $2.01 \times 10^{-7}$ | $2.67 \times 10^{-8}$ | $3.33 \times 10^{-5}$ | $0.003$ | $0.003$ | $0.003$ | $0.607$ | $0.607$ |
| $I_n$     | LMI-based [8] | 1.8456 | 0.8549 | 0.2917 | 0.876 | 0.895 | 0.968 | 0.870 | 0.913 |
|           | Inverse KKT [10] | 1.013* | 1.056* | 5.361* | 5.173* | 1.092* | 5.277* | 1.016* | 5.306* |
| Proposed  | $3.22 \times 10^{-6}$ | $6.28 \times 10^{-9}$ | $1.05 \times 10^{-6}$ | $0.003$ | $0.003$ | $0.003$ | $0.607$ | $0.607$ |
| $10I_n$   | LMI-based [8] | 2.38 | 2.294 | 0.919 | 1.59 | 0.999 | 0.933 | 0.934 | 0.973 |
|           | Inverse KKT [10] | 1.013* | 1.061* | 1.167* | 5.168* | 1.1389 | 3.278* | 1.0194* | 5.141* |

* Invalid Solution Estimates: the matrices $\hat{Q}$ and $\hat{R}$ have negative eigenvalues.

TABLE 2. A comparison of error $||\Lambda - \hat{\Lambda}||_2$ of the model-free algorithm (Algorithm 1) and the model-based algorithm (Algorithm 2) for a Monte-Carlo simulation results (500 episodes).

| $\Sigma_w$ | Method | $T_j = 2$ | $T_j = 2$ | $T_j = 2$ | $T_j = 50$ | $T_j = 50$ | $T_j = 50$ | $T_j = 50$ | $T_j = 50$ |
|------------|--------|------------|------------|------------|------------|------------|------------|------------|------------|
|           | Algorithm 1 | 0.400 | 0.4638 | 0.3566 | $0.0039$ | 0.0014 | 0.0013 | $0.005$ | 0.0031 |
|           | Algorithm 2 | 0.081 | 0.0032 | 0.0019 | 0.004 | 0.0008 | 0.0006 | 0.0005 | 0.0003 |
|           | Algorithm 1 | 0.7351 | 0.6101 | 0.5428 | $0.6853$ | 0.2607 | $0.1114$ | 0.6064 | $0.2017$ |
|           | Algorithm 2 | 1.085 | 0.8187 | $0.5339$ | 0.7385 | $0.2482$ | 0.1493 | $0.5926$ | 0.2298 |
| Proposed  | Algorithm 1 | 0.6085 | 0.6837 | $0.7073$ | $0.8412$ | $0.8599$ | 0.5928 | $0.7872$ | 0.8987 |
|           | Algorithm 2 | 1.2622 | 1.1237 | 1.028 | 1.1749 | 0.9270 | $0.4897$ | 1.1297 | $0.8164$ |

FIGURE 1. The Monte-Carlo simulation results (500 times) using the proposed Model-Free Stochastic IOC LQR algorithm for the noisy case.

FIGURE 2. The length of trajectories $k_{\Sigma}$ method in [10] is only guaranteed to obtain valid results for the cases in which there is a large data set with low noise and $(\hat{A}, \hat{B}) = (A, B)$ perfectly known. To ensure well-posedness, the method in [8] artificially restricts the manifold of possible solutions to guarantee the uniqueness of any obtained solution. As such, our proposed method outperforms that of [8] by virtue of having a larger searchable solution space, while still guaranteeing well-posedness. Additionally, the method in [8] does not explicitly model the disturbance in the development of their IOC LQR methodology. As such, this lack of accounting may be the reason why their method performs badly for disturbances with larger covariances. We also note that in general, when the estimated dynamics $(\hat{A}, \hat{B})$ is ill-fitted, the performance of all the methods, including our proposed method, worsens. This is to be expected as the model estimation error affects the computation of $\Lambda$.

D. COMPARING MODEL-FREE TO MODEL-BASED PERFORMANCES

In this last section, we compare the performance results of Algorithm 1 to that of the Algorithm 2, as per Table 2. We see that the model-free algorithm (Algorithm 1) obtains an overall better robust performance to noise, while the model-based algorithm (Algorithm 2), generally, yields better results

the covariance of the noise, and the estimates of the system matrices $(\hat{A}, \hat{B})$. For example, if there are too few data points, then the rank of the recovery matrix is too low; or if the covariance of the noise is large, then the rank of the recovery is too high. Table 1 shows that the inverse KKT
when the covariance $\Sigma_{rr}$ is small, and when the number of trajectories $r$ is comparable to or larger than the length of the trajectories. These results are congruous with our insights from Theorem 3.

VI. CONCLUSION

In this paper, we proposed a Model-Free Stochastic Inverse Optimal Control (IOC) linear quadratic (LQ) control algorithm for determining the optimal control gain and cost function parameters from partial or incomplete noisy system trajectories. This technique surpasses known IOC LQ and IRL techniques by providing low complexity bounds for reverse-engineering black-box LQ optimal controllers from multiple partial state and input trajectories in a model-free manner. An immediate extension to this work is to consider the Stochastic Model-Free IOC LQ with non-Gaussian noise.

APPENDIX A EXTENSION TO THE IOC LQ TRACKING (LQT) PROBLEM

The discrete-time infinite-horizon stochastic IOC LQT problem seeks to minimize a cost function:

$$J_T \doteq \lim_{T \to \infty} \frac{1}{T} \mathbb{E} \left[ \sum_{k=0}^{T-1} e_k^T Q e_k + u_k^T R u_k \right], \quad (24)$$

where the system dynamics in (1) is required to track a desired reference signal $r_k$ as closely as possible. The reference trajectory propagates with the dynamics $r_{k+1} = A_r r_k$ where $A_r \in \mathbb{R}^{n \times n}$ is a stable system matrix, i.e., $\rho(A_r) < 1$. The error signal $e_k \triangleq x_k - r_k$ has the dynamics:

$$e_{k+1} = A e_k + B u_k + w_k - A_r r_k \quad (25)$$

This inspires us to consider the augmented state vector $z_k \in \mathbb{R}^{2n} \triangleq [x_k^T, e_k^T]^T$ with the dynamics:

$$z_{k+1} = \begin{bmatrix} A & 0 \\ -A_r & A_r \end{bmatrix} z_k + \begin{bmatrix} B \\ B \end{bmatrix} u_k + \begin{bmatrix} I \\ 0 \end{bmatrix} w_k, \quad (26)$$

and the equivalent cost function,

$$J_T = \lim_{T \to \infty} \frac{1}{T} \mathbb{E} \left[ \sum_{k=0}^{T-1} z_k^T Q z_k + u_k^T R u_k \right], \quad (27)$$

where $Q_z \in \mathbb{R}^{2n \times 2n} \triangleq \text{blkdiag}(0, Q)$. The LQT problem is now in LQR form and can be efficiently solved using Algorithm 1.

APPENDIX B PROOF OF THEOREM 3

In order to prove Theorem 3, we use the following supporting Lemma. Let a matrix $M = (M_{jk}) \in \mathbb{R}^{d \times d}$ and $\|M\|_{\max} \doteq \max_{j \leq k} |M_{jk}|$.

Lemma 5. [37, Proof of Lemma 1, pg. 3141] Let $S$ be the marginal sample covariance matrix defined as

$$S \doteq \frac{1}{T} \sum_{t=1}^T X_t X_t^T \quad (28)$$

for a stationary lag 1 autoregressive process $X_t \in \mathbb{R}^d$ generated by

$$X_t = A_1 X_{t-1} + Z_t, \quad Z_t \sim \mathcal{N}(0, \Sigma), \quad (29)$$

then, we have

$$\mathbb{P} (\|S - \Sigma\|_{\max} > \eta) \leq 3d^2 \exp \left( \frac{T}{2} \right) + 3d^2 \exp \left( -\frac{T}{2} \left( \frac{\eta \min \{\lambda_{ji}(\Sigma_{ij}) (1 - |A_i|)\} - 2T^{-\frac{3}{2}} \} \right)^2 \right) \quad (30)$$

where $\Sigma \succ 0 \in \mathbb{R}^{d \times d}$ is the covariance of $X_t$.

Now, we prove the result in Theorem 3.

Proof of Theorem 3. Let $S_j \doteq \frac{1}{T_j + 1} \sum_{k=0}^{T_j} v_{jk}^T v_{jk}^T$ so that

$$\|X^0_r - X\|_{\max} = \left\| \frac{1}{r} \sum_{j=1}^r S_j - X \right\|_{\max} \leq \left\| \frac{1}{r} \sum_{j=1}^r (S_j - X) \right\|_{\max} \leq \left\| \frac{1}{r} \sum_{j=1}^r \min_{i \leq j} \|X_{i}\|_{\max} \right\|_{\max} \leq 1 \left\| \frac{1}{r} \sum_{j=1}^r 16 \|X_{i}\|_{\max} \right\|_{\max} \quad (31)$$

Using Lemma 5, we have

$$\mathbb{P} (\|S_t - X\|_{\max} > \eta, \cdots, \|S_r - X\|_{\max} > \eta) \leq \sum_{j=1}^r \left[ 3(n + m)^2 \exp \left( -\frac{T_j + 1}{2} \right) + 3(n + m)^2 \exp \left( -\frac{T_j + 1}{2} + (T_j + 1)^2 \right) \right] \quad (32)$$

This concludes the proof of (18).

ACKNOWLEDGMENT

The authors would like to acknowledge that this work is supported by NSF CNS-1836952.

REFERENCES

[1] R. E. Kalman, “When Is a Linear Control System Optimal?” Journal of Basic Engineering, vol. 86, no. 1, pp. 51–60, 3 1964.
[2] B. D. Anderson, “The inverse problem of optimal control.” Stanford Uni. Calif. Stanford Electronics Labs, Tech. Rep. TR-6560-3, 1966.
[3] A. McNeill, Optima for animals. Princeton University Press, 2021.
[4] P. Petousis, A. Winter, W. Speier, D. R. Aberle, W. Hsu, and A. A. T. Bui, “Using sequential decision making to improve lung cancer screening performance,” IEEE Access, 2019.
[5] B. Molinari, “The stable regulator problem and its inverse,” IEEE Transactions on Automatic Control, vol. 18, no. 5, pp. 454–459, 1973.
[6] T. Fuji and M. Narazaki, “A Complete Optimality Condition in the Inverse Problem of Optimal Control,” SIAM journal on control and optimization, vol. 22, no. 2, pp. 327–341, 1984.
[7] N. Ab Azar, A. Shahmansoorian, and M. Davoudi, “From inverse optimal control to inverse reinforcement learning: A historical review,” Annual Reviews in Control, vol. 50, pp. 119–138, 2020.
[8] M. C. Priess, R. Conway, J. Choi, J. M. Popovich, and C. Radcliffe, “Solutions to the inverse lqr problem with application to biological systems analysis,” IEEE Transactions on Control Systems Technology, vol. 23, no. 2, pp. 770–777, 2015.

[9] H. Zhang, J. Umenberger, and X. Hu, “Inverse optimal control for discrete-time finite-horizon Linear Quadratic Regulators,” Automatica, vol. 110, p. 108593, 2019.

[10] W. Jin, D. Kulić, S. Mou, and S. Hirche, “Inverse optimal control from incomplete trajectory observations,” The International Journal of Robotics Research, vol. 40, no. 6-7, pp. 848–865, 2021.

[11] T. L. Molloy, J. J. Ford, and T. Perez, “Online inverse optimal control for control-constrained discrete-time systems on finite and infinite horizons,” Automatica, vol. 120, p. 109109, 2020.

[12] C. Yu, Y. Li, H. Fang, and J. Chen, “System identification approach for inverse optimal control of finite-horizon linear quadratic regulators,” Automatica, vol. 129, p. 109636, 2021.

[13] P. Abbeel and A. Y. Ng, “Apprenticeship learning via Inverse Reinforcement Learning,” in Proceedings of the twenty-first international conference on Machine learning, 2004, p. 1.

[14] S. Levine and V. Koltun, “Continuous inverse optimal control with locally optimal examples,” in Proceedings of the 29th International Conference on International Conference on Machine Learning. Omnipress, 2012, p. 475–482.

[15] J. Choi and K.-E. Kim, “Hierarchical Bayesian Inverse Reinforcement Learning,” IEEE transactions on cybernetics, vol. 45, no. 4, pp. 793–805, 2014.

[16] W. Xue et al., “Inverse reinforcement learning in tracking control based on inverse optimal control,” IEEE Transactions on Cybernetics, pp. 1–12, 2021.

[17] I. Clavera et al., “Model-based Reinforcement Learning via Meta-Policy Optimization,” in Conference on Robot Learning. PMLR, 2018, pp. 617–629.

[18] D. P. Bertsekas et al., Dynamic programming and optimal control 3rd edition, volume II. Athena scientific Belmont, 2011.

[19] C. De Persis and P. Tesi, “Low-complexity learning of Linear Quadratic Regulators from noisy data,” Automatica, vol. 128, p. 109548, 2021.

[20] F. Orkan and Y. Ma, “Modeling Driver Behavior in Car-Following Interactions With Automated and Human-Driven Vehicles and Energy Efficiency Evaluation,” IEEE Access, vol. 9, 2021.

[21] S. Arora and P. Doshi, “A survey of inverse reinforcement learning: Challenges, methods and progress,” Artificial Intelligence, vol. 297, p. 103500, 2021.

[22] A. Komanduru and J. Honorio, “On the correctness and sample complexity of inverse reinforcement learning,” Advances in Neural Information Processing Systems, vol. 32, pp. 7112–7121, 2019.

[23] A Komanduru and J. Honorio, “A Lower Bound for the Sample Complexity of Inverse Reinforcement Learning,” arXiv preprint arXiv:2103.04446, 2021.

[24] D. P. Bertsekas et al., Dynamic programming and optimal control: Vol. 1. Athena scientific Belmont, 2000.

[25] D. Lee and I. Hu, “Primal-Dual Q-learning Framework for LQR Design,” IEEE Transactions on Automatic Control, vol. 64, no. 9, pp. 3756–3763, 2019.

[26] H. J. van Waarde et al., “Willems’ Fundamental Lemma for State-Space Systems and Its Extension to Multiple Datasets,” IEEE Control Systems Letters, vol. 4, no. 3, pp. 602–607, 2020.

[27] I. M. Jan C. Willems, Paolo Rapisarda and B. De Moor, “A Note on Persistency of Excitation,” Systems and Control Letters, vol. 54, no. 4, pp. 325–329, 2005.

[28] F. Wang, S. Mukherjee, S. Richardson, and S. M. Hill, “High-dimensional regression in practice: an empirical study of finite-sample prediction, variable selection and ranking,” Statistics and computing, vol. 30, no. 3, pp. 697–719, 2020.

[29] F. E. Harrell et al., Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. Springer, 2001, vol. 608.

[30] S. Sastry and M. Bodson, Adaptive control: stability, convergence and robustness. Courier Corporation, 2011.

[31] H. Mania, S. Tu, and B. Recht, “Certainty equivalence is efficient for linear quadratic control,” arXiv preprint arXiv:1902.07826, 2019.

[32] G. Gu, Discrete-Time Linear Systems: Theory and Design with Applications. Springer Science & Business Media, 2012.

[33] L. Vandenberghe and S. Boyd, “Semidefinite programming,” SIAM review, vol. 38, no. 1, pp. 49–95, 1996.
SHANELLE G. CLARKE received the Bachelor’s degree in Aerospace Engineering from the Massachusetts Institute of Technology, MA, USA in 2015. She received the MSc in Control Systems degree from Imperial College, London, UK in 2016. She is currently pursuing a PhD from the School of Aeronautics and Astronautics, Purdue University. Her research interests include high-performance nonlinear control, adaptive robust control, reinforcement learning control and data-driven optimal control methods. With respect to application, her interests include flight navigation and control for UASs in mixed cooperative-competitive environments.

SOOYUNG BYEON received the bachelor’s and master’s degrees in aerospace engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea, in 2012 and 2014, respectively. He was a senior research engineer at Satrix Initiative Co., Ltd., Daejeon, Republic of Korea, from 2014 to 2019. He is currently a research assistant with the School of Aeronautics and Astronautics, Purdue University, West Lafayette, Indiana, USA. His research interests include human-automation interaction, human behavior modeling, imitation learning, and applied experiments. He was awarded with Ross Fellowship presented by the College of Engineering, Purdue University, in 2019.

INSEOK HWANG received the Ph.D. degree in Aeronautics and Astronautics from Stanford University, and is currently a professor in the School of Aeronautics and Astronautics and a university faculty scholar at Purdue University. His research interests lie in high assurance autonomy (which guarantees safety, security and performance) for Cyber-Physical Systems (CPS) based on the hybrid systems approach and its applications to safety critical systems such as aircraft/spacecraft/unmanned aerial systems (UAS), air traffic management, and multi-agent systems, which are complex systems with interacting cyber (or logical or human) elements and physical components. For his research, he leads the Flight Dynamics and Control/Hybrid Systems Laboratory at Purdue University. He received various awards, including the National Science Foundation (NSF) CAREER award in 2008, recognition as one of the nation’s brightest young engineers by the National Academy of Engineering (NAE) in 2008, the AIAA Special Service Citation in 2010, the University Faculty Scholar by Purdue university in recognition of outstanding scholarship in 2017, Outstanding Graduate Faculty Mentor award 2018, and the C.T. Sun Award in recognition of excellence in research in 2019. He is an Associate Fellow of AIAA and a member of IEEE Control Systems Society and Aerospace and Electronic Systems Society. He is currently an associate editor of the IEEE Transactions on Aerospace and Electronic Systems.