2024 European Control Conference (ECC)
June 25-28, 2024. Stockholm, Sweden

Combining Learning and Control in Linear Systems

Andreas A. Malikopoulos, *IEEE Senior Member*

***Abstract*— In this paper, we provide a theoretical framework that separates the control and learning tasks in a linear system. This separation allows us to combine offline model-based control with online learning approaches and thus circumvent current challenges in deriving optimal control strategies in applications where a large volume of data is added to the system gradually in real time and not altogether in advance. We provide an analytical example to illustrate the framework.**

I. Introduction

REINFORCEMENT learning (RL) [1], [2] has emerged as an adaptive method to control systems [3] with unknown dynamics [4]. There have also been research efforts on developing learning approaches using Bayesian analysis to address such problems [5]. Other approaches over the years have focused on direct or indirect RL methods including robust learning-based [6], [7], learning-based model predictive control [8]–[10] on autonomous racing cars [11], real-time learning [12], [13] of powertain systems with respect to the driver's driving style [14], [15], learning for traffic control [16] for transferring optimal policies [17], [18], decentralized learning for stochastic games [19], learning for optimal social routing [20] and congestion games [21], and learning for enhanced security against replay attacks in cyber-physical systems [22].

The implications of the strategies derived using a model, which is typically different from the actual system, have been reported in [23]. A recent paper [24] proposed approximate learning of an information state to address problems when the dynamics of the actual system are not known. Other efforts have combined adaptive control with RL to derive control strategies in real time [25]. Space constraints prevent us from discussing the complete list of papers reported in the literature in this area. Two survey papers [26], [27], however, include a comprehensive review of the RL approaches.

In some applications, we encounter a volume of data gradually incorporated into the system. To derive the optimal control strategy in such applications, we typically use a model [28]. However, model-based control might not effectively facilitate optimal solutions partly due to the existing discrepancy between the model and the actual system. On the other hand, supervised learning approaches might not always facilitate robust solutions using training data derived offline. Similarly, RL approaches might impose undesired implications on the system's robustness.

In this paper, we investigate how to circumvent these challenges at the intersection of learning and control. We derive sufficient statistics that can represent the system's growing data. This sufficient statistics is called *information state* of the system and takes values in a time-invariant space. This information state can be used to derive *separated control strategies*, which are related to the separation between the estimation of the information state and the derivation of the control strategy. Given this separation, for any control strategy at time t, the evolution of the information state of the system does not depend on the control strategy at t but only on the realization value of the control at t [29]. Thus, the evolution of the information states is separated from the choice of the current control strategy. Hence, the optimal control strategy is derived offline using the information state, which can be learned online using standard techniques [30], [31] while data are incorporated into the system. This approach departs from traditional model-based and supervised (or unsupervised) learning approaches. The framework could effectively facilitate optimal solutions in a wide range of applications where a large volume of data is added to the system gradually in real time and not altogether in advance, such as emerging mobility systems, mobility markets, smart power grids, power systems, social media platforms, robot cooperation, and the Internet of Things.

The structure of the paper is organized as follows. In Section II, we present the formulation of the optimal control problem. In Section III, we introduce the separated control strategies. In Section IV, we illustrate the framework with a simple analytical example. Finally, we draw concluding remarks in Section V.

II. Problem Formulation

A. Notation

In our exposition, we denote by $\mathbb{E}[\cdot]$ the expectation of random variables, by $\mathbb{P}(\cdot)$ the probability of an event, and by $p(\cdot)$ the probability density function. We denote by $\mathbb{E}^{\mathbf{g}}[\cdot]$, $\mathbb{P}^{\mathbf{g}}(\cdot)$, and $p^{\mathbf{g}}(\cdot)$ that the expectation, probability, and probability density function, respectively, depending on the choice of the control strategy $\mathbf{g}$. Random variables are denoted with upper case letters, and their realizations with lower case letters, e.g., for a random variable X_t, x_t denotes its realization. In some occasions, we denote the expected value of a random variable with lower case letter, e.g., $\mathbb{E}[X_t] = x_t$. Subscripts denote time. The shorthand notation $X_{0:T}$ denotes the the vector of random variables $(X_0, \dots, X_T)$, and the shorthand notation $x_{0:T}$ denotes the vector of their realization $(x_0, \dots, x_T)$.

This work was supported by NSF under Grants CNS-2149520 and CMMI-2219761.

The author is with the School of Civil and Environmental Engineering, Cornell University, Ithaca, NY 14850, USA. (email: amaliko@cornell.edu.)

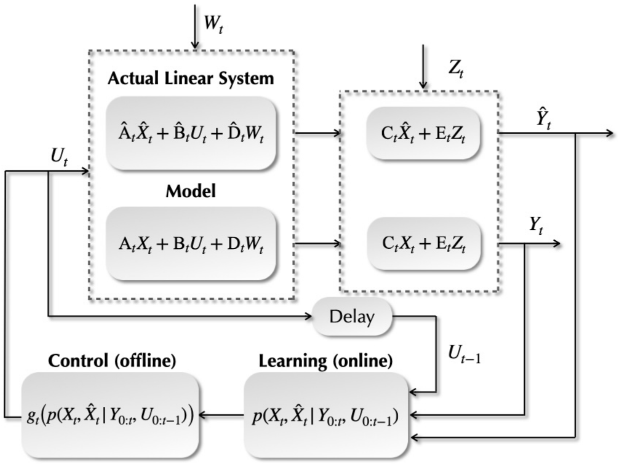
W_t
Actual Linear System
$\hat{A}_t\hat{X}_t + \hat{B}_tU_t + \hat{D}_tW_t$
Model
$A_tX_t + B_tU_t + D_tW_t$
Z_t
$C_t\hat{X}_t + E_tZ_t$
$C_tX_t + E_tZ_t$
$\hat{Y}_t$
Y_t
U_t
Delay
U_{t-1}
Control (offline)
Learning (online)
$g_t(p(X_t, \hat{X}_t \mid Y_{0:t}, U_{0:t-1}))$
$p(X_t, \hat{X}_t \mid Y_{0:t}, U_{0:t-1})$

Fig. 1: The proposed framework on separating learning and control, where the separated control strategy is applied to both the actual system and the system's model in parallel.

B. Modeling Framework

We consider a linear system in which a volume of data is added to the system gradually and not altogether in advance. We aim to find sufficient statistics that can be used to compress the increasing data of the system without loss of optimality [32]. These statistics are a conditional probability of the system's state at time $t \in \mathbb{R}_{\geq 0}$ given all data available up until t, which is called the information state of the system. We use this information state to derive separated control strategies. By deriving separated control strategies, we can derive the optimal control strategy offline with respect to the information state and then use learning methods to learn the information state online.

In particular, in our framework illustrated in Fig. 1, we seek to use the actual linear system that we wish to optimally control online, in parallel with a model of the system that is available. Let X_t, $t = 0, 1, \ldots, T$, $T \in \mathbb{N}$, be a random variable that corresponds to the state of the system's model and $\hat{X}_t$, $t = 0, 1, \ldots, T$, be a random variable that corresponds to the state of the actual system. Both X_t and $\hat{X}_t$ are defined on an appropriate probability space and take values in $\mathbb{R}^n$, $n \in \mathbb{N}$. The control U_t of the actual system is a random variable defined on the same probability space and takes values in $\mathbb{R}^m$, $m \in \mathbb{N}$. Given an initial state X_0, the model of the linear system is

$$X_{t+1} = \mathbf{A}_t X_t + \mathbf{B}_t U_t + \mathbf{D}_t W_t, \;\; t = 0, 1, \ldots, T-1, \quad (1)$$

where $\mathbf{A}_t, \mathbf{B}_t$, and $\mathbf{D}_t$ are matrices of appropriate dimensions, and $W_t \in \mathbb{R}^r$, $r \in \mathbb{N}$, is a random variable that corresponds to the external, uncontrollable disturbance. Given the same initial state X_0, the actual system is represented by

$$\hat{X}_{t+1} = \hat{\mathbf{A}}_t \hat{X}_t + \hat{\mathbf{B}}_t U_t + \hat{\mathbf{D}}_t W_t, \;\; t = 0, 1, \ldots, T-1, \quad (2)$$

where $\hat{\mathbf{A}}_t, \hat{\mathbf{B}}_t$, and $\hat{\mathbf{D}}_t$ are matrices of appropriate dimensions. The sequence $\{W_t;\; t = 0, 1, \ldots, T-1\}$ is a sequence of independent random variables independent of the initial state X_0.

At the time $t = 0, 1, \ldots, T-1$, we make an observation $Y_t \in \mathbb{R}^p$, $p \in \mathbb{N}$, of the model's output described by the observation equation

$$Y_t = \mathbf{C}_t X_t + \mathbf{E}_t Z_t, \quad (3)$$

where $\mathbf{C}_t, \mathbf{E}_t$ are matrices of appropriate dimensions, and $Z_t \in \mathbb{R}^s$, $s \in \mathbb{N}$, is a random variable that represents the sensor's noise. Similarly, at time t, we make an observation $\hat{Y}_t \in \mathbb{R}^p$, $p \in \mathbb{N}$, of the actual system, described by the observation equation

$$\hat{Y}_t = \mathbf{C}_t \hat{X}_t + \mathbf{E}_t Z_t, \quad (4)$$

Note $\{Z_t\}$, $t = 0, \ldots, T-1$, is a sequence of independent random variables that are also independent of $\{W_t\}$, $t = 0, \ldots, T-1$, and the initial state X_0.

A control strategy $\mathbf{g} = \{g_t\}$ of the system yields a decision

$$U_t = g_t(\hat{Y}_{0:t}, U_{0:t-1}), \;\; t = 0, \ldots, T-1, \quad (5)$$

where the measurable function g_t is the control law. The feasible set of the control strategies is $\mathcal{G}$, i.e., $\mathbf{g} \in \mathcal{G}$.

Problem 1 [Actual linear system]: Derive the optimal control strategy $\mathbf{g}^* \in \mathcal{G}$ which minimizes the following cost of the actual system,

$$\hat{J}(\mathbf{g}) = \mathbb{E}^{\mathbf{g}} \left[\sum_{t=0}^{T-1} c_t(\hat{X}_t, U_t) + c_T(\hat{X}_T) \right], \quad (6)$$

where the expectation in (6) is taken with respect to the joint probability distribution of $\hat{X}_t$ and U_t imposed by the control strategy $\mathbf{g} \in \mathcal{G}$; $c_t(\cdot,\cdot) : \mathcal{X} \times \mathcal{U}_t \to \mathbb{R}$ is the cost function of the system at t, and $c_T(\cdot) : \mathcal{X} \to \mathbb{R}$ is the cost function at T. The probability distribution of the primitive random variables X_0, $\{W_t\}$, $\{Z_t\}$, the cost functions $\{c_t(\cdot,\cdot)\}$ for $t = 0, \ldots, T-1$ and $c_T(\cdot)$, and the matrices $\mathbf{C}_t, \mathbf{E}_t$ for $t = 0, \ldots, T-1$ are all known. However, the matrices $\hat{\mathbf{A}}_t, \hat{\mathbf{B}}_t, \hat{\mathbf{D}}_t$ are not known for $t = 0, \ldots, T-1$.

III. Separating Learning and Control Tasks

Let $\mathbf{g} = \{g_t;\; t = 0, \ldots, T-1\}$, $\mathbf{g} \in \mathcal{G}$, be a control strategy which yields a decision $U_t = g_t(Y_{0:t}, U_{0:t-1})$ using the model of the linear system. We establish an information state by using in parallel the system's model and the actual system as shown in Fig. 1.

The probability density function $p(X_t, \hat{X}_t \mid Y_{0:t}, U_{0:t-1})$ is the information state (defined formally next), denoted by $\Pi_t(Y_{0:t}, U_{0:t-1})(X_t, \hat{X}_t)$. To simplify notation, in what follows, the information state $\Pi_t(Y_{0:t}, U_{0:t-1})$ $(X_t, \hat{X}_t)$ at t is denoted by Π_t without the arguments, which will be used only if it is required in our exposition.

Definition 1. *The information state,* $\Pi_t(Y_{0:t}, U_{0:t-1})$ $(X_t, \hat{X}_t)$*, is (a) a function of* $(Y_{0:t}, U_{0:t-1})$*, and (b) its evolution* $\Pi_{t+1}(Y_{0:t+1}, U_{0:t})(X_{t+1}, \hat{X}_{t+1})$ *at the next step* $t+1$ *depends on* $\Pi_t(Y_{0:t}, U_{0:t-1})(X_t, \hat{X}_t)$*,* Y_{t+1}*, and* U_t*.*

Theorem 1. *The information state $\Pi_t(Y_{0:t}, U_{0:t-1})(X_t, \hat{X}_t)$ does not depend on the control strategy $\mathbf{g} \in \mathcal{G}$. Furthermore, there exists a function ϕ_t such that*

$$\begin{aligned}&\Pi_{t+1}(Y_{0:t+1}, U_{0:t})(X_{t+1}, \hat{X}_{t+1}) \\ &= \phi_t\big[\Pi_t(Y_{0:t}, U_{0:t-1})(X_t, \hat{X}_t), Y_{t+1}, U_t\big],\end{aligned} \tag{7}$$

for all $t = 0, 1, \ldots, T-1$.

The result of Theorem 1 is equivalent to the result of [33, Theorem 2] when the system's information structure is classical [34], [35] and the controller has perfect recall [29], [36].

Definition 2. *A control strategy $\mathbf{g} \in \mathcal{G}$, $\mathbf{g} = \{g_t\}$, $t = 0, \ldots, T-1$, is called separated if g_t depends on $Y_{0:t} = (Y_0, \ldots, Y_t)$ and $U_{0:t-1} = (U_0, \ldots, U_{t-1})$ through the information state, i.e., $U_t = g_t\big(\Pi_t(Y_{0:t}, U_{0:t-1})(X_t, \hat{X}_t)\big)$. Let $\mathcal{G}^s \subseteq \mathcal{G}$ denote the set of all separated control strategies.*

Since the dynamics of the actual system are not known, we cannot solve Problem 1. Thus, to obtain the optimal strategy in Problem 1, we formulate the following problem that we solve offline using the system's model (1).

Problem 2: Derive offline the optimal separated strategy $\mathbf{g}^* \in \mathcal{G}^s$ to minimize the following cost function

$$J(\mathbf{g}; \hat{x}_{0:T}) = \mathbb{E}^{\mathbf{g}}\left[\sum_{t=0}^{T-1}\Big[c_t(X_t, U_t) + \beta \cdot |Y_{t+1} - \hat{Y}_{t+1}|^2\Big] + c_T(X_T)\right],$$

or, using (3) and (4),

$$J(\mathbf{g}; \hat{x}_{0:T}) = \mathbb{E}^{\mathbf{g}}\left[\sum_{t=0}^{T-1}\Big[c_t(X_t, U_t) + \beta \cdot |X_{t+1} - \hat{X}_{t+1}|^2\Big] + c_T(X_T)\right], \tag{8}$$

where β adjusts the units of the norm accordingly, while the norm penalizes the discrepancy between the expected values of the state of the system's model and the state of the actual system. Since we solve (8) offline using model (1), no information about the actual system is available, and thus the expected values $\hat{x}_{0:T} = (\hat{x}_0, \ldots, \hat{x}_T)$ of the states $\hat{X}_{0:T} = (\hat{X}_0, \ldots, \hat{X}_T)$ of the actual system are not known. Hence, when we derive the optimal control strategy $\mathbf{g}^*$, it is parameterized with respect to all possible values $\hat{x}_{0:T}$.

Next, to obtain offline the optimal separated control strategy in Problem 2, we use the information state $\Pi_t(Y_{0:t}, U_{0:t-1})(X_t, \hat{X}_t)$. It can be shown [33] that we can derive a classical dynamic program decomposition with respect to Π_t to yield a separated control strategy, namely, a control strategy $\mathbf{g} = \{g_t\}$, $t = 0, \ldots, T-1$ where g_t depends on $Y_{0:t+1}$ and $U_{0:t}$ only through the information state, i.e., $U_t = g_t\big(\Pi_t(Y_{0:t}, U_{0:t-1})(X_t, \hat{X}_t)\big)$.

The separated control strategy is derived offline, thus, it is parameterized with respect to the potential expected values $\hat{x}_{0:T}$ of the state $\hat{X}_t$ of the actual system. Then, we apply the parameterized strategy to the actual system and the system's model in parallel (Fig. 1), and we collect data from both. Using these data, we compute $\Pi_t(Y_{0:T}, U_{0:T-1})(X_{t+1}, \hat{X}_{t+1})$ online.

Proposition 1. *The information state $\Pi_t(Y_{0:t}, U_{0:t-1})(X_t, \hat{X}_t)$ of the system illustrated in Fig. 1 can be represented as a function of $p(X_t \mid Y_{0:t}, U_{0:t-1})$, $p(\hat{X}_t \mid \hat{Y}_{0:t}, U_{0:t-1})$, and $p(\hat{Y}_{0:t} \mid U_{0:t-1})$.*

Proof. Recall

$$\Pi_t(Y_{0:t}, U_{0:t-1})(X_t, \hat{X}_t) = p(X_t, \hat{X}_t \mid Y_{0:t}, U_{0:t-1}). \tag{9}$$

Next,

$$\begin{aligned}&p(X_t, \hat{X}_t \mid Y_{0:t}, U_{0:t-1}) \\ &= \frac{p(\hat{X}_t \mid X_t, Y_{0:t}, U_{0:t-1}) \cdot p(X_t, Y_{0:t}, U_{0:t-1})}{p(Y_{0:t}, U_{0:t-1})} \\ &= \frac{p(\hat{X}_t \mid U_{0:t-1}) \cdot p(X_t, Y_{0:t}, U_{0:t-1})}{p(Y_{0:t}, U_{0:t-1})} \\ &= p(\hat{X}_t \mid U_{0:t-1}) \cdot p(X_t \mid Y_{0:t}, U_{0:t-1}).\end{aligned} \tag{10}$$

In the second equality, we used the fact that $\hat{X}_t$ does not depend on X_t and $Y_{0:t}$, and in the third equality, we applied Bayes' rule.

Next, we write the first term in (10) as follows

$$p(\hat{X}_t \mid U_{0:t-1}) = \sum_{\hat{Y}_{0:t}} p(\hat{X}_t \mid \hat{Y}_{0:t}, U_{0:t-1}) \cdot p(\hat{Y}_{0:t} \mid U_{0:t-1}). \tag{11}$$

Substituting (11) into (10), the result follows. □

Remark 1. *The conditional probability $p(X_t \mid Y_{0:t}, U_{0:t-1})$ can be obtained easily using the model offline. The conditional probability $p(\hat{X}_t \mid \hat{Y}_{0:t}, U_{0:t-1})$ can be obtained from the Kalman filter to estimate $\hat{X}_t$ first, and then through recursive equations starting from the initial prior $p(\hat{X}_0 \mid \hat{Y}_0, U_0)$. The conditional probability $p(\hat{Y}_{0:t} \mid U_{0:t-1})$ can be obtained using standard approaches [30], [31]. Ongoing research focuses on enhancing our understanding of the computational implications in learning $p(\hat{Y}_{0:t} \mid U_{0:t-1})$ in real time.*

As we operate both the actual system and the model using the separated control strategy (Fig. 1), we compute $p(\hat{X}_t \mid \hat{Y}_{0:t}, U_{0:t-1})$ and learn $p(\hat{Y}_{0:t} \mid U_{0:t-1})$ that allows us to compute the information state $\Pi_t(Y_{0:t}, U_{0:t-1})(X_t, \hat{X}_t)$ (the conditional probability $(p(X_t \mid Y_{0:t}, U_{0:t-1})$ is known a priori from the model). Next, we show that when the information state $\Pi_t(Y_{0:t}, U_{0:t-1})(X_t, \hat{X}_t)$ becomes known, then the separated control strategy is optimal for the actual system.

Theorem 2. *Let $\mathbf{g} \in \mathcal{G}^s$ be an optimal separated control strategy parameterized with respect to $\hat{x}_{0:T}$, derived offline using the system's model, that minimizes the following cost* *function,*

$$J(\mathbf{g}; \hat{x}_{0:T}) := \mathbb{E}^{g}\left[\sum_{t=0}^{T-1}\Big[c_t(X_t, U_t) + \beta \cdot |X_{t+1} - \hat{X}_{t+1}|^2\Big] \right.$$
$$\left. + c_T(X_T)\right], \tag{12}$$

in Problem 2. Then, if $p(X_t, \hat{X}_t \mid Y_{0:t}, U_{0:t-1})$ *becomes known, then* $\mathbf{g}$ *is also optimal for Problem 1,*

$$\hat{J}(\mathbf{g}) = \mathbb{E}^{g}\left[\sum_{t=0}^{T-1} c_t(\hat{X}_t, U_t) + c_T(\hat{X}_T)\right]. \tag{13}$$

Proof. Suppose that the minimum value of the cost function $c_T(X_T)$ occurs at $X_T = x_T \in \mathbb{R}^n$. Hence,

$$c_T(X_T = x_T) = c_T(\hat{X}_T = x_T). \tag{14}$$

Suppose that the minimum value of the cost function $c_t(X_t, U_t)$ at $t = 0, \ldots, T-1$ occurs at $X_t = x_t \in \mathbb{R}^n$ and corresponds to the optimal control $U_t = u_t^*$. Then, the minimum value in the one-time-step cost in (12) at $t = 0, \ldots, T-1$ is when the expected value of the cost function is $c_t(x_t, u_t^*)$ and $\mathbb{E}[|X_{t+1} - \hat{X}_{t+1}|^2] = 0$, hence

$$\min_{u_t} \mathbb{E}^{\mathbf{g}}\Big[c_t(X_t, u_t) + \beta \cdot |X_{t+1} - \hat{X}_{t+1}|^2\Big] = c_t(x_t, u_t^*). \tag{15}$$

Since at each time $t = 0, \ldots, T-1$, the separated control strategy $\mathbf{g} \in \mathcal{G}^s$ yields a control input $u_t' = g_t\big(p(x_t, \hat{x}_t \mid y_{0:t}, u_{0:t-1})\big)$ such that

$$\begin{aligned} u_t' &= \arg\min_{u_t} \mathbb{E}\Big[c_t(X_t, u_t) + \beta \cdot |X_{t+1} - \hat{X}_{t+1}|^2\Big] \\ &= \arg\min_{u_t} c_t(x_t, u_t^*), \end{aligned} \tag{16}$$

this implies that $u_t' = u_t^*$.

By summing up all minimum expected values of the cost function $c_t(\cdot, \cdot)$ at each $t = 0, \ldots, T-1$ and $c_T(\cdot)$ at $t = T$ corresponding to $\mathbf{g} \in \mathcal{G}^s$, we obtain (13). □

IV. Illustrative Example

In this section, we present a simple example of deriving the optimal control strategy for a linear system by separating the learning and control tasks. The purpose of the example is to demonstrate in simple steps the proposed framework. The primitive random variables, i.e., the initial state, X_0, and disturbance, W_0, of the system, are Gaussian random variables with zero mean, variance 1, and covariance 0.5. The state of the actual system is denoted by $\hat{X}_t$, $t = 0, 1, 2$. The evolution of the system is described by the following equations

$$\begin{aligned} \hat{X}_0 &= X_0, \\ \hat{X}_1 &= \hat{X}_0 + U_0 + W_0 = X_0 + U_0 + W_0, \\ \hat{X}_2 &= \hat{X}_1 + U_1. \end{aligned} \tag{17}$$

We assume that we have a complete observation of the state, i.e.,

$$\hat{Y}_t = \hat{X}_t, \quad t = 0, 1, 2. \tag{18}$$

A control strategy $\mathbf{g} = \{g_t;\ t = 0, 1\}$, $\mathbf{g} \in \mathcal{G}$, where g_t is the control law, yields the control action U_t, $t = 0, 1$, of the system, i.e.,

$$U_0 = g_0(\hat{Y}_0) = g_0(\hat{X}_0) = g_0(X_0), \tag{19}$$

$$U_1 = g_1(\hat{Y}_{0:1}, U_0) = g_1(\hat{X}_{0:1}, U_0) = g_1(X_0, \hat{X}_1, U_0). \tag{20}$$

We seek to derive the optimal control strategy $\mathbf{g}^* \in \mathcal{G}$ of the system represented in (17) which minimizes the following expected cost:

$$J(\mathbf{g}) = \min_{u_0 \in \mathcal{U}_0, u_1 \in \mathcal{U}_1} \frac{1}{2}\mathbb{E}^{\mathbf{g}}\left[(\hat{X}_2)^2 + (U_1)^2\right]. \tag{21}$$

We pretend that the dynamics of the system in (17) (the actual system) are not known, but we have available the following model that can be used to obtain $\mathbf{g} \in \mathcal{G}$:

$$\begin{aligned} X_0 &= X_0, \\ X_1 &= 3X_0 + 2U_0 + 2W_0, \\ X_2 &= 3X_1 + 3U_1, \end{aligned} \tag{22}$$

with

$$Y_t = X_t, \ t = 0, 1, 2. \tag{23}$$

From (17) and (22), we note that there exists a discrepancy between the actual system and the model that is available.

A. Optimal Control Strategy

First, we obtain the optimal control strategy $\mathbf{g}^* \in \mathcal{G}$ of the actual system using (17).

The cost for the actual system (17) is

$$\begin{aligned} J(\mathbf{g}) &= \min_{u_0 \in \mathcal{U}_0, u_1 \in \mathcal{U}_1} \frac{1}{2}\mathbb{E}^{\mathbf{g}}\left[(\hat{X}_2)^2 + (U_1)^2\right] \\ &= \min_{u_0 \in \mathcal{U}_0, u_1 \in \mathcal{U}_1} \frac{1}{2}\mathbb{E}^{\mathbf{g}}\left[(\hat{X}_1 + U_1)^2 + (U_1)^2\right] \\ &= \min_{u_0 \in \mathcal{U}_0, u_1 \in \mathcal{U}_1} \frac{1}{2}\mathbb{E}^{\mathbf{g}}\left[(X_0 + U_0 + W_0 + U_1)^2 + (U_1)^2\right]. \end{aligned} \tag{24}$$

If the dynamics of the actual system given in (17) were known, then we could use (24) to derive the optimal control strategy $\mathbf{g}^* \in \mathcal{G}$. Since the primitive random variables are Gaussian with zero mean, variance 1, and covariance 0.5, we can use the linear least-squares estimator to compute the unique optimal solution of (24), which is

$$U_0 = -\frac{1}{2}X_0, \quad U_1 = -\frac{1}{4}X_0 - \frac{1}{2}W_0. \tag{25}$$

B. Solution Through Separation Between Learning and Control

In this section, we consider that the dynamics of the actual system (17) are not known, but we have the model (22) of the system available. Using this model, we can obtain the optimal control strategy by applying the framework presented in Section III. More specifically, we use (22) and seek to derive the separated control strategy $\mathbf{g} \in \mathcal{G}^s$, $\mathbf{g} = \{g_t;\ t = 0, 1\}$, where the control laws are $g_0\big(\mathbb{P}(X_0, \hat{X}_0 \mid Y_0)\big)$ and $g_1\big(\mathbb{P}(X_1, \hat{X}_1 \mid Y_0, Y_1, U_0)\big)$, that minimizes the following cost (see Theorem 2),

$$\begin{aligned} &J(\mathbf{g}; \hat{x}_{0:2}) \\ &= \min_{u_0 \in \mathcal{U}_0, u_1 \in \mathcal{U}_1} \frac{1}{2} \mathbb{E}^{\mathbf{g}}\big[(X_2)^2 + (U_1)^2 \\ &+ \beta(X_1 - \hat{X}_1)^2 + \beta(X_2 - \hat{X}_2)^2) \mid X_0, X_1, U_0\big]. \end{aligned} \tag{26}$$

From (22) and taking $\beta = 1$, (26) becomes

$$\begin{aligned} &J(\mathbf{g}; \hat{x}_{0:2}) \\ &= \min_{u_0 \in \mathcal{U}_0, u_1 \in \mathcal{U}_1} \frac{1}{2} \mathbb{E}^{\mathbf{g}}\Big[(3X_1 + 3U_1)^2 + (U_1)^2 \\ &+ (X_1 - \hat{X}_1)^2 + (X_2 - \hat{X}_2)^2) \mid X_0, X_1, U_0\Big] \\ &= \min_{u_0 \in \mathcal{U}_0, u_1 \in \mathcal{U}_1} \frac{1}{2} \mathbb{E}^{\mathbf{g}}\Big[\big(3(3X_0 + 2U_0 + 2W_0) + 3U_1\big)^2 \\ &+ (U_1)^2 + (X_1 - \hat{X}_1)^2 + (X_2 - \hat{X}_2)^2) \mid X_0, X_1, U_0\Big]. \end{aligned} \tag{27}$$

The cost in (27) becomes equal to the original cost in (24), if the control action U_0 and U_1 make the last two terms equal to zero, i.e.,

$$\mathbb{E}^{\mathbf{g}}[X_1 - \hat{X}_1] = \mathbb{E}^{\mathbf{g}}[3X_0 + 2U_0 + 2W_0 - \hat{X}_1 \mid X_0] = 0, \tag{28}$$

$$\mathbb{E}^{\mathbf{g}}[X_2 - \hat{X}_2] = \mathbb{E}^{\mathbf{g}}[3X_1 + 3U_1 - \hat{X}_2 \mid X_0, X_1, U_0] = 0. \tag{29}$$

From (28), it follows that

$$\begin{aligned} \mathbb{E}^{\mathbf{g}}[U_0] &= \mathbb{E}^{\mathbf{g}}\Big[\frac{\hat{X}_1 - 3X_0 - 2W_0}{2} \mid X_0\Big] \\ &= g_0\big(p(X_0, \hat{X}_0 \mid X_0)\big). \end{aligned} \tag{30}$$

Similarly, from (29), it follows that

$$\begin{aligned} \mathbb{E}^{\mathbf{g}}[U_1] &= \mathbb{E}^{\mathbf{g}}\Big[\frac{\hat{X}_2 - 3X_1}{3} \mid X_0, X_1, U_0\Big] \\ &= \mathbb{E}^{\mathbf{g}}\Big[\frac{\hat{X}_2 - 3(3X_0 + 2U_0 + 2W_0)}{3} \mid X_0, X_1, U_0\Big] \\ &= \mathbb{E}^{\mathbf{g}}\Big[\frac{\hat{X}_2 - 9X_0 - 6U_0 - 6W_0}{3} \mid X_0, X_1, U_0\Big] \\ &= g_1\big(p(X_1, \hat{X}_1 \mid X_0, X_1, U_0)\big). \end{aligned} \tag{31}$$

Thus, U_0 and U_1 in (30) and (31), respectively, are parameterized with respect to the expected values of the state of the actual system, i.e., $\hat{x}_0 = x_0$, $\hat{x}_1$ and $\hat{x}_2$, and make the last two terms in (27) vanish.

Next, we use the control actions U_0 and U_1 derived by the separated control strategies $g_0\big(p(X_0, \hat{X}_0 \mid X_0)\big)$ and $g_1\big(p(X_1, \hat{X}_1 \mid X_0, X_1, U_0)\big)$ in (30) and (31), respectively, to control both the actual linear system (17) and the model (22) (see Fig. 1). As we operate both systems, we compute the information states $p(X_0, \hat{X}_0 \mid X_0)$ and $p(X_1, \hat{X}_1 \mid X_0, X_1, U_0)$. However, from Proposition 1, we know that to compute $p(X_0, \hat{X}_0 \mid X_0)$ and $p(X_1, \hat{X}_1 \mid X_0, X_1, U_0)$, it is sufficient to compute the conditional probabilities $p(X_0 \mid X_0)$, $p(X_1 \mid X_0, X_1, U_0)$, and $p(\hat{X}_0 \mid \hat{X}_0, \hat{X}_1, U_0)$, and to learn $p(\hat{X}_0, \hat{X}_1 \mid U_0, U_1)$. Once we compute these conditional probabilities, the expected values of U_0 and U_1 in (30) and (31) become known.

By substituting (30) in (27), we obtain

$$\begin{aligned} J(\mathbf{g}; \hat{x}_{0:2}) &= \min_{u_0 \in \mathcal{U}_0, u_1 \in \mathcal{U}_1} \frac{1}{2} \mathbb{E}^{\mathbf{g}}\Big[\Big(3(3X_0 + 2\frac{\hat{X}_1 - 3X_0 - 2W_0}{2} \\ &\quad + 2W_0) + 3U_1\Big)^2 + (U_1)^2 \mid X_0, X_1, U_0\Big] \\ &= \min_{u_0 \in \mathcal{U}_0, u_1 \in \mathcal{U}_1} \frac{1}{2} \mathbb{E}^{\mathbf{g}}\Big[\big(3\hat{X}_1 + 3U_1\big)^2 + (U_1)^2 \\ &\quad \mid X_0, X_1, U_0\Big] \\ &= \min_{u_0 \in \mathcal{U}_0, u_1 \in \mathcal{U}_1} \frac{1}{2} \mathbb{E}^{\mathbf{g}}\Big[\big(3(X_0 + U_0 + W_0) + 3U_1\big)^2 \\ &\quad + (U_1)^2 \mid X_0, X_1, U_0\Big]. \end{aligned} \tag{32}$$

However, at $t = 0$, we do not consider U_1 and X_1. Thus, the last equation becomes

$$\min_{u_0 \in \mathcal{U}_0} \frac{1}{2} \mathbb{E}^{\mathbf{g}}\Big[\big(3(X_0 + U_0 + W_0)\big)^2 \mid X_0\Big]. \tag{33}$$

The optimization problem above is to choose for each value X_0 the best estimate, in a mean squared error sense, of $(X_0 + U_0 + W_0)$, which yields $U_0 = -\frac{1}{2}X_0$ which is the same solution as in (25). By substituting (31) into the model (22), we obtain

$$\begin{aligned} X_2 &= 3X_1 + 3\frac{\hat{X}_2 - 9X_0 - 6U_0 - 6W_0}{3} \\ &= 3(3X_0 + 2U_0 + 2W_0) + \hat{X}_2 - 9X_0 - 6U_0 - 6W_0 \\ &= \hat{X}_2, \end{aligned} \tag{34}$$

hence the expected total cost $J(\mathbf{g}; \hat{x}_{0:2})$ in (26) becomes

$$\begin{aligned} J(\mathbf{g}; \hat{x}_{0:2}) &= \min_{u_0 \in \mathcal{U}_0, u_1 \in \mathcal{U}_1} \frac{1}{2} \mathbb{E}^{\mathbf{g}}\Big[(\hat{X}_2)^2 + (U_1)^2)\Big] \\ &= \min_{u_0 \in \mathcal{U}_0, u_1 \in \mathcal{U}_1} \frac{1}{2} \mathbb{E}^{\mathbf{g}}\Big[(\hat{X}_1 + U_1)^2 + (U_1)^2)\Big]. \end{aligned} \tag{35}$$

The minimum in (35) at time $t = 1$ can be found by taking the partial derivative with respect to U_1 which yields

$$\begin{aligned} &\mathbb{E}^{\mathbf{g}}\Big[\big(\hat{X}_1 + U_1 + U_1\big)\Big] = \mathbb{E}^{\mathbf{g}}\Big[\big(X_0 + U_0 + W_0 + U_1 + U_1\big)\Big] \\ &= 0 \end{aligned} \tag{36}$$

that results in the same solution $U_1 = -\frac{1}{4}X_0 - \frac{1}{2}W_0$ as in (25).

V. Concluding Remarks

In this paper, we presented a theoretical framework that provides a data-driven approach for linear systems at the intersection of learning and control. The framework separates the control and learning tasks which allows us to combine offline model-based control with online learning approaches and thus circumvent current challenges in deriving optimal control strategies. One feature that distinguishes the framework presented here from other learning-based or combined learning and control approaches reported in the literature is that the large volume of data added to the system is compressed to sufficient statistics, without loss of optimality, that takes values in a time-invariant space. Hence, as the volume of data added to the systems increases, the domain of the control strategies does not increase with time. Ongoing research investigates the computational implications of learning the information state. In our exposition, we restricted attention to centralized control systems. A potential future research direction includes expanding the framework to decentralized systems [37].

References

[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Bradford Books, 1998.

[2] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*. Athena Scientific, 1996.

[3] R. S. Sutton, A. G. Barto, and R. J. Williams, "Reinforcement learning is direct adaptive optimal control," *IEEE Control Systems Magazine*, vol. 12, no. 2, pp. 19–22, 1992.

[4] A. B. Kordabad, D. Reinhardt, A. S. Anand, and S. Gros, "Reinforcement learning for mpc: Fundamentals and current challenges," *IFAC-PapersOnLine*, vol. 56, no. 2, pp. 5773–5780, 2023.

[5] J. F. Fisac, A. K. Akametalu, M. N. Zeilinger, S. Kaynama, J. Gillula, and C. J. Tomlin, "A general safety framework for learning-based control in uncertain robotic systems," *IEEE Transactions on Automatic Control*, vol. 64, no. 7, pp. 2737–2752, 2019.

[6] P. Bouffard, A. Aswani, and C. Tomlin, "Learning-based model predictive control on a quadrotor: Onboard implementation and experimental results," in *2012 IEEE International Conference on Robotics and Automation*, 2012, pp. 279–284.

[7] A. Aswani, H. Gonzalez, S. S. Sastry, and C. Tomlin, "Provably safe and robust learning-based model predictive control," *Automatica*, vol. 49, no. 5, pp. 1216–1226, 2013.

[8] X. Zhang, M. Bujarbaruah, and F. Borrelli, "Near-optimal rapid mpc using neural networks: A primal-dual policy learning framework," *IEEE Transactions on Control Systems Technology*, pp. 1–13, 2020.

[9] U. Rosolia and F. Borrelli, "Learning model predictive control for iterative tasks. a data-driven control framework," *IEEE Transactions on Automatic Control*, vol. 63, no. 7, pp. 1883–1896, 2018.

[10] L. Hewing, K. P. Wabersich, M. Menner, and M. N. Zeilinger, "Learning-based model predictive control: Toward safe learning in control," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 3, no. 1, pp. 269–296, 2021/05/31 2020.

[11] U. Rosolia and F. Borrelli, "Learning how to autonomously race a car: A predictive control approach," *IEEE Transactions on Control Systems Technology*, vol. 28, no. 6, pp. 2713–2719, 2020.

[12] A. A. Malikopoulos, P. Y. Papalambros, and D. N. Assanis, "A real-time computational learning model for sequential decision-making problems under uncertainty," *Journal of Dynamic Systems, Measurement and Control, Transactions of the ASME*, vol. 131, no. 4, 2009.

[13] A. A. Malikopoulos, *Real-Time, Self-Learning Identification and Stochastic Optimal Control of Advanced Powertrain Systems*. ProQuest, 2011.

[14] A. A. Malikopoulos, P. Y. Papalambros, and D. N. Assanis, "Online identification and stochastic control for autonomous internal combustion engines," *Journal of Dynamic Systems, Measurement, and Control*, vol. 132, no. 2, pp. 024 504–024 504, 2010.

[15] A. A. Malikopoulos, D. N. Assanis, and P. Y. Papalambros, "Real-time self-learning optimization of diesel engine calibration," *Journal of Engineering for Gas Turbines and Power*, vol. 131, no. 2, 2009.

[16] C. Wu, A. Kreidieh, K. Parvate, E. Vinitsky, and A. Bayen, "Flow: Architecture and benchmarking for reinforcement learning in traffic control," *IEEE Transations on Robotics, TRO-17-0544*, 2017.

[17] B. Chalaki, L. E. Beaver, B. Remer, K. Jang, E. Vinitsky, A. Bayen, and A. A. Malikopoulos, "Zero-shot autonomous vehicle policy transfer: From simulation to real-world via adversarial learning," in *IEEE 16th International Conference on Control & Automation (ICCA)*, 2020, pp. 35–40.

[18] K. Jang, E. Vinitsky, B. Chalaki, B. Remer, L. Beaver, A. A. Malikopoulos, and A. Bayen, "Simulation to scaled city: zero-shot policy transfer for traffic control via autonomous vehicles," in *Proceedings of the 10th ACM/IEEE International Conference on Cyber-Physical Systems*, 2019, pp. 291–300.

[19] G. Arslan and S. Yüksel, "Decentralized q-learning for stochastic teams and games," *IEEE Transactions on Automatic Control*, vol. 62, no. 4, pp. 1545–1558, 2017.

[20] W. Krichene, M. S. Castillo, and A. Bayen, "On social optimal routing under selfish learning," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 1, pp. 479–488, 2018.

[21] W. Krichene, B. Drighès, and A. M. Bayen, "Online learning of nash equilibria in congestion games," *SIAM Journal on Control and Optimization*, vol. 53, no. 2, pp. 1056–1081, 2015.

[22] P. P. Sahoo and K. G. Vamvoudakis, "On-off adversarially robust q-learning," *IEEE Control Systems Letters*, vol. 4, no. 3, pp. 749–754, 2020.

[23] A. D. Kara and S. Yüksel, "Robustness to incorrect system models in stochastic control and application to data-driven learning," in *2018 IEEE Conference on Decision and Control (CDC)*, 2018, pp. 2753–2758.

[24] J. Subramanian, A. Sinha, R. Seraj, and A. Mahajan, "Approximate information state for approximate planning and reinforcement learning in partially observed systems," *Journal of Machine Learning Research*, vol. 23, pp. 1–83, 2022.

[25] A. Guha and A. Annaswamy, "Online policies for real-time control using mrac-rl," *ArXiv*, vol. abs/2103.16551, 2021.

[26] B. Recht, "A tour of reinforcement learning: The view from continuous control," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 2, pp. 253–279, 2019.

[27] B. Kiumarsi, K. G. Vamvoudakis, H. Modares, and F. L. Lewis, "Optimal and autonomous control using reinforcement learning: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2042–2062, 2018.

[28] A. A. Malikopoulos, "A duality framework for stochastic optimal control of complex systems," *IEEE Transactions on Automatic Control*, vol. 61, no. 10, pp. 2756–2765, 2016.

[29] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification and Adaptive Control*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1986.

[30] L. Gyorfi and M. Kohler, "Nonparametric estimation of conditional distributions," *IEEE Transactions on Information Theory*, vol. 53, no. 5, pp. 1872–1879, 2007.

[31] M. Brand, "Structure learning in conditional probability models via an entropic prior and parameter extinction," *Neural Computation*, vol. 11, no. 5, pp. 1155–1182, 1999.

[32] C. Striebel, "Sufficient statistics in the optimum control of stochastic systems," *Journal of Mathematical Analysis and Applications*, vol. 12, no. 3, pp. 576–592, 1965.

[33] A. A. Malikopoulos, "Separation of learning and control for cyber-physical systems," *Automatica*, vol. 151, no. 110912, 2023.

[34] S. Yüksel and T. Basar, *Stochastic Networked Control Systems*, 2013th, Ed. Birkhäuser, 2013.

[35] J. H. van Schuppen and T. Villa, *Coordination Control of Distributed Systems*. Springer, 2015.

[36] D. P. Bertsekas, D. P. Bertsekas, D. P. Bertsekas, and D. P. Bertsekas, *Dynamic programming and optimal control*. Athena scientific Belmont, MA, 1995, vol. 1, no. 2.

[37] A. A. Malikopoulos, "On team decision problems with nonclassical information structures," *IEEE Transactions on Automatic Control*, vol. 68, no. 7, pp. 3915–3930, 2023.