Information-theoretic multi-time-scale partially observable systems with inspiration from leukemia treatment

Margaret P. Chapman a, Emily Jensen b, Steven M. Chan c, Laurent Lessard d

a Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, 10 King’s College Road, Toronto, ON, M5S 3G8, Canada

b Department of Electrical, Computer and Energy Engineering, University of Colorado Boulder, 425 UCB, Boulder, CO, 80309, United States

c Department of Medical Biophysics, Faculty of Medicine, University of Toronto, Princess Margaret Cancer Research Tower, MaRS Centre, 101 College Street, Toronto, ON, M5G 1L7, Canada

d Department of Mechanical and Industrial Engineering, Northeastern University, 334 Snell Engineering Center, 360 Huntington Avenue, Boston, MA, 02115, United States

Abstract

We study a partially observable nonlinear stochastic system with unknown parameters, where the given time scales of the states and measurements may be distinct. The proposed setting is inspired by disease management, particularly leukemia.

Key words: Stochastic control; Nonlinear systems; Partially observable systems; System identification; Biomedical systems.

1 Introduction

A patient with a disease may be viewed as a stochastic system that evolves in response to control inputs, e.g., drug doses. Aspects of the patient can be measured, e.g., a tumor can be scanned or a blood sample can be taken. Controls and measurements typically operate on different time scales; for example, measurements may be weekly, and there may be days when no drug is taken.

While there is knowledge about the underlying biochemical processes, such processes may be nonlinear, noisy, and may vary between patients. In many medical settings (e.g., leukemia treatment, diabetes treatment, and blood pressure management), the doses of drugs require adjustments for the purpose of regulation (e.g., keeping white blood cells, glucose, and blood pressure, respectively, within specific ranges). From a mathematical point of view, it is unclear how to unify the above properties (partial observability, nonlinear stochastic dynamics with uncertain parameters, and multiple time scales) into a rigorous control-theoretic framework. Our aim is to develop such a unifying framework.

We consider a discrete-time partially observable stochastic system that differs from a standard set-up. The system has an unknown deterministic parameter vector θ, representing interpatient variability. Furthermore, the states and measurements may operate on different time scales, the dynamics and measurement equations may be nonlinear, the process and measurement noise may have unbounded support, and the spaces of states, controls, and measurements do not have finite cardinality. Working in the setting above, we aim to develop a pathway that provides a control policy with awareness about the uncertainty in an estimate for θ and sensitivity of the state relative to θ. A theoretical foundation must be developed before approximations can be investigated to permit computation.

Our contribution is of a conceptual nature. We offer a theoretical framework that unifies the following key characteristics: partial observability, uncertain parameters, multiple time scales, and nonlinear stochastic dynamics (Fig. 1). First, we devise a theoretical approach to handle the unknown parameters, building on concepts from optimum experiment design and dual control (Sec. 3). Second, we represent the system as a multi-time-scale partially observable (PO) Markov decision process (MDP) with information-theoretic cost functions and show that it enjoys regularity properties under some assumptions (Sec. 4). Third, we provide conditions that guarantee the existence of an optimal policy for a belief-space MDP corresponding to the POMDP (Sec. 5). Our main contribution is to reduce the complications that...
arise from the unobservable state, unknown parameters, multiple time scales, and nonlinear stochastic dynamics to a form that admits a mathematical solution. We exemplify the assumptions in the context of a leukemia treatment model in Section 6. Although our presented framework does not solve the curse of dimensionality, it does provide a launching point for bridging the gap between theory and practice, as further described in our conclusion (Sec. 7).

1.1 Related literature

The theory is related to three bodies of literature: POMDPs, optimum experiment design, and stochastic adaptive and dual control. POMDPs with Borel state and control spaces have been studied since at least the 1970s [31] [36] [4, Ch. 10] [13, Ch. 4]. The state-of-the-art conditions that facilitate policy synthesis for these general POMDPs were developed recently [9] [21] [10]. We propose and study a POMDP problem with unknown parameters and information-theoretic costs, where the time scales of the measurements and states may be distinct. The different time scales in the setting of unknown parameters pose difficulties due to the irregular updates to the parameter estimates and posterior distributions. We offer a reformulation that uses measure-theoretic first principles and regularity conditions proposed by [9] [21]. The author of [13, Ch. 4.5] studies POMDPs with unknown parameters, continuous bounded cost functions, and weakly continuous bounded dynamics for posterior distributions. However, we consider cost functions that are lower semicontinuous and bounded below. We do not assume weakly continuous bounded dynamics for posterior distributions in view of a counterexample [9, Ex. 4.2]. We propose stage costs that penalize poor quality of information from measurements using the Fisher information matrix from optimum experiment design (to be discussed). To accommodate these information-theoretic costs, we assume Euclidean spaces of states, parameters, measurements, and measurement noise, and we make assumptions about the existence and continuity of Jacobians.

An optimum experiment design (OED) problem is a type of optimal control problem [32]. An example OED problem is to optimize the planned measurement times by maximizing the quality of information, which can be quantified using the Fisher information matrix (FIM) [29] [32] [28] [8]. Then, an optimal control problem can be solved using the parameter estimate [32, Eq. (3.3)] [19, Algorithm 2, pp. 79–81]. In clinical practice, the planned measurement times are determined by the patient’s schedule, which may be infeasible to adjust. Thus, we are not concerned with optimizing the planned measurement times. The OED literature includes measurement noise but often neglects process noise [32] [24] [19] [8]; an exception is [33]. We include both measurement and process noise; the latter plays numerous roles in biochemical systems [30] [7]. The techniques in [33] rely on linear approximations and a Riccati differential equation to approximate the covariance of the estimated state at the final time. In contrast, we leverage the theory of discrete-time nonlinear POMDPs for policy synthesis. Our objective evaluates the stage-wise performance of the states and controls and an FIM-based criterion.

This paper has connections to stochastic adaptive and dual control, which are related to OED. Separating the problems of parameter estimation and controller design characterizes classical adaptive control schemes, e.g., self-tuning regulators [2, p. 22]. In general, separating the two problems implies that parameter uncertainties are neglected in controller synthesis, and hence, dual control methods have been developed to tackle the two problems simultaneously [2, pp. 22–24]. A recent review of stochastic dual control is provided by [27], and a concise summary is provided by [15, p. 276]. We discuss some recent papers on dual control, which involve systems with unknown parameters (exception: [11]). Model predictive controllers for linear systems with process noise [3] and nonlinear systems without process noise [24] have been developed, where the objective is a sum of a performance metric and an FIM-based metric. The authors of [24] also assume initial state distributions with bounded support. Heirung et al. develop a model predictive controller using a quadratically-constrained quadratic program to minimize the predicted mean-squared output error; the system is a linear regression model with normally distributed noise and parameters [12]. Model predictive controllers with information matrix-based constraints have been proposed for linear systems with process noise [25] and nonlinear systems without process noise [34]. The controller in [25] incorporates different time scales for the measurements and controls. Feng and Houska develop a real-time model predictive control algorithm for a nonlinear system to optimize a sum of a performance metric and an approximation for the average loss of optimality due to poor future state estimates [11]. They assume that the measurement and process noise have bounded support to facilitate the latter approximation, which employs an extended Kalman filter [11]. In contrast to the above papers, we provide a theoretical study of a partially observable multi-time-scale nonlinear system with unknown parameters, where the measurement and process noise may have unbounded support. We include an FIM-based criterion in the objective rather than as a constraint to avoid potential feasibility issues.

1.2 Notation

\( \mathbb{R} \) is the real line, \( \mathbb{R}^\infty := \mathbb{R} \cup \{-\infty, +\infty\} \) is the extended real line, \( \mathbb{R}^n \) is \( n \)-dimensional Euclidean space,
\( \mathbb{R}^+_n := \{ x \in \mathbb{R}^n : x_i > 0, i = 1, 2, \ldots, n \} \), and \( \mathbb{N} \) is the set of natural numbers. If \( f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m \) with \( (x, y) \mapsto f(x, y) \), then \( \frac{\partial f(a, b)}{\partial x} \in \mathbb{R}^{m \times n} \) is the Jacobian matrix of \( f \) with respect to \( x \) evaluated at \( (a, b) \in \mathbb{R}^n \times \mathbb{R}^l \). The Euclidean norm is \( \| \cdot \| \). If \( A \in \mathbb{R}^{m \times n} \), we define \( (\ast)^T A := A^T \). We define \( \varphi_x : \mathbb{R}^l \to \mathbb{R}^l \) by \( \varphi_x(x) := (\max\{x_1, \epsilon\}, \max\{x_2, \epsilon\}, \ldots, \max\{x_1, \epsilon\})^T \), where \( \epsilon \in \mathbb{R} \) is near zero. If \( M \) is a separable metrizable space, then \( B_M \) is the Borel sigma algebra on \( M \), \( P_M \) is the collection of probability measures on \( (M, B_M) \) with the weak topology, and \( \| \|_{TV} \) is the total variation norm on \( P_M \). If \( p_i \in P_M \) for \( i = 1, 2 \), the total variation norm of \( p_1 - p_2 \) is \( \| p_1 - p_2 \|_{TV} = 2 \sup_{M \in B_M} | p_1(M) - p_2(M) | \). We use the notation \( P \) to denote a Borel set, i.e., \( \in \mathcal{B}_M \), following the notation from [4]. If \( y \in M \), then \( \delta_y \) is the Dirac measure in \( M \), \( \mathbb{R} \), then \( \sigma(x) \) is defined by \( \sigma(x) := (\max\{x_1, \epsilon\}, \max\{x_2, \epsilon\}, \ldots, \max\{x_1, \epsilon\})^T \), where \( \epsilon \in \mathbb{R} \) is near zero. If \( M \) is a separable metrizable space, then \( B_M \) is the Borel sigma algebra on \( M \), \( P_M \) is the collection of probability measures on \( (M, B_M) \) with the weak topology, and \( \| \|_{TV} \) is the total variation norm on \( P_M \). If \( p_i \in P_M \) for \( i = 1, 2 \), the total variation norm of \( p_1 - p_2 \) is \( \| p_1 - p_2 \|_{TV} = 2 \sup_{M \in B_M} | p_1(M) - p_2(M) | \).

2 System description

We consider a discrete-time stochastic system that differs from a standard one. There is an unknown deterministic parameter vector, there may be different time scales, and the state is not observable. The system takes the following form:

\[
\begin{align*}
x_{t+1} &= f_t(x_t, u_t, d_t; \theta), \quad t \in \mathbb{T}, \\
y_t &= h_t(x_t; \theta) + w_t, \quad t \in \mathbb{T}_y,
\end{align*}
\]

where \( x_t \in \mathcal{S} \) is a state, \( y_t \in \mathcal{Y} := \mathbb{R}^m \) is a measurement, \( u_t \in \mathcal{U} \) is a control, \( d_t \in \mathcal{D} \) is a process noise realization, \( w_t \in \mathcal{W} \) is a measurement noise realization, and \( \theta \in \mathcal{P} := \mathbb{R}^p \) is a parameter vector whose true value is unknown. An initial estimate for \( \theta \) is available. The state space \( \mathcal{S} \subseteq \mathbb{R}^n \) is a nonempty Euclidean space. The control space \( \mathcal{C} \) and the process noise space \( \mathcal{D} \) are nonempty Borel spaces [4, Def. 7.7]. The states evolve on \( \mathbb{T} := \{0, 1, \ldots, N - 1\} \), a time horizon of length \( N \in \mathbb{N} \). The measurements evolve on a nonempty subset \( \mathbb{T}_y \) of \( \mathbb{T} \). The controls may be optimized on a nonempty subset \( \mathbb{T}_u \) of \( \mathbb{T} \). If \( t \in \mathbb{T} \setminus \mathbb{T}_u \), then \( u_t \) is a given default value \( u \in \mathcal{U} \). The functions \( f_t : \mathcal{S} \times \mathcal{C} \times \mathcal{D} \times \mathcal{P} \to \mathcal{S} \) for every \( t \in \mathbb{T} \) and \( h_t : \mathcal{S} \times \mathcal{P} \to \mathcal{Y} \) for every \( t \in \mathbb{T}_y \) are measurable. These functions and the horizons \( \mathbb{T}, \mathbb{T}_y, \) and \( \mathbb{T}_u \) are given. The quantities \( x_t, u_t, d_t, y_t, \) and \( w_t \) in (1) are realizations of random objects \( X_t, U_t, D_t, Y_t, \) and \( W_t \), respectively. The random objects \( X_0, D_t \) for every \( t \in \mathbb{T} \), and \( W_t \) for every \( t \in \mathbb{T}_y \) are independent. \( D_t \) for any \( t \in \mathbb{T} \) and \( W_t \) for any \( t \in \mathbb{T}_y \) do not depend on \( \theta \). The (prior) distribution of \( X_0 \) is given, and the distributions of \( D_t \) and \( W_t \) are given when these random objects are defined. (Assuming knowledge of such distributions is typical in research about partially observable systems.)

We use the following notations:

- If \( t \in \mathbb{T} \setminus \mathbb{T}_y \), then the distribution of \( D_t \) is \( \mu_t \in \mathcal{P}_D \).
- If \( t \in \mathbb{T}_y \), then the distribution of \( W_t \) is \( \rho_t \in \mathcal{P}_W \).
- If \( t \in \mathbb{T}_y \setminus \{N\} \), then the distribution of \( (D_t, W_t) \) is \( \nu_t \in \mathcal{P}_D \times \mathcal{Y} \).

\( \rho_t \) and \( \nu_t \) are related by \( \rho_t(\mathcal{Y}) = \nu_t(\mathcal{D} \times \mathcal{Y}) \) for every \( \mathcal{Y} \in \mathcal{B}_Y \) and \( t \in \mathbb{T}_y \setminus \{N\} \). The system (1) has a stage cost function \( c_t : \mathcal{C} \times \mathcal{D} \times \mathcal{Y} \to \mathbb{R}^+ \) for every \( t \in \mathbb{T} \) and a terminal cost function \( c_N : \mathcal{C} \times \mathcal{P} \to \mathbb{R}^+ \), which may depend on \( \theta \). These functions assess the performance of the states and controls. We invoke the following assumption.

Assumption 1 (About \( f_t, c_t, \mathcal{C}, h_t, \rho_t \)) We assume:

- (a) \( f_t \) is continuous for every \( t \in \mathbb{T} \);
- (b) \( c_t \) is l.s.c. and bounded below for every \( t \in \mathbb{T}_N \);
- (c) \( \mathcal{C} \) is compact;
- (d) \( h_t \) is continuous for every \( t \in \mathbb{T}_y \);
3 Information-theoretic control system

Toward designing a control policy for (1), here we provide a pathway for estimating \( \theta \) and managing potential differences between \( T_x \) and \( T_y \). First, we specify cost functions to assess the quality of measurements using concepts from optimum experiment design. Second, we define an enlarged system to record these information-theoretic costs. We conclude the section by studying the regularity of observation kernels. Part (f) will facilitate estimating \( \theta \) using measurements and assessing the quality of the measurements. In particular, our leukemia treatment model satisfies A1 (Sec. 6).

3.1 Formulating information-theoretic costs

We would like to extract useful information from measurements to inform the estimation of \( \theta \). The OED literature assesses the usefulness of information through the Fisher information matrix (FIM). For the system (1) and a state trajectory \((x_0, x_1, \ldots, x_N) \in \mathbb{S}^{N+1}\), the FIM \( \mathcal{F} \in \mathbb{R}^{p \times p} \) is defined by \( \mathcal{F} := \sum_{t \in \mathbb{T}_y} \mathcal{F}_t((x_t^\top, \xi_t^\top, \theta_t^\top)^\top) \) [32, Def. 3.2], where each term \( \mathcal{F}_t((x_t^\top, \xi_t^\top, \theta_t^\top)^\top) := (s)^\top \left( \frac{\partial h_t(x_t; \theta)}{\partial \theta} \right), \quad t \in \mathbb{T}_y, \quad (2a) \)

depends on the output sensitivity matrix [23, Eq. (6.101)]
\[
\frac{\partial h_t(x_t; \theta)}{\partial \theta} := \frac{\partial h_t(x_t; \theta)}{\partial x} \xi_t + \frac{\partial h_t(x_t; \theta)}{\partial \theta}, \quad t \in \mathbb{T}_y, \quad (2b)
\]
and the state sensitivity matrix \( \xi_t := \frac{dx_t}{d\theta} \in \mathbb{R}^{n \times p} \). The output sensitivity matrix (2b) is the total derivative of the predicted measurement \( h_t(x_t; \theta) \) with respect to \( \theta \). The \( \xi_t \)-dynamics are defined by [1]

\[
(\xi_{t+1} = \phi_t(\xi_t, x_t, u_t, d_t; \theta), \quad t \in \mathbb{T}, \quad (3a)
\]

where
\[
\phi_t(x, u, d; \theta) := \frac{\partial f_t(x, u, d; \theta)}{\partial x} \cdot \xi + \frac{\partial f_t(x, u, d; \theta)}{\partial \theta}.
\]

If \( \theta \) does not include the initial state \( x_0 \), then \( \xi_0 \) is the zero matrix in \( \mathbb{R}^{n \times p} \). If \( \theta \) does include the initial state, then \( \xi_0 \) consists of an identity matrix and a zero matrix.

We would like the size of the FIM \( \mathcal{F} \), e.g., its trace or determinant, to be large [32, Def. 3.4]. Statistical background underlying \( \mathcal{F} \) can be found in [26]. We content ourselves with an intuitive explanation. A large trace(\( \mathcal{F} \)) means that the predicted measurement \( h_t(x_t; \theta) \) varies a large amount with respect to \( \theta \). Therefore, different values of \( \theta \) correspond to different measurements, which facilitates the estimation of the true \( \theta \) from the measurements. Defining an augmented state that includes the state sensitivity matrix \( \xi_t \) is useful for assessing FIM-based criteria in optimal control problems [25].

The equations (2)–(3) depend on \( \theta \), whose true value is unknown. Hence, one evaluates these equations using an estimate for \( \theta \), which we denote by \( \hat{\theta}_t \in \mathbb{P} \). Consequently, we must track the evolution of the estimates, which is an aspect of dual control [2, p. 23] [13, Ch. 2.5].

In view of the above, we consider an augmented state \( \chi_t := (x_t^\top, \xi_t^\top, \theta_t^\top)^\top \) with values in \( \mathbb{X} := \mathbb{S} \times \mathbb{R}^{n \times p} \times \mathbb{P} \). We specify a cost function \( \tilde{c}_t : \mathbb{X} \to \mathbb{R} \), which aims to penalize a small trace of \( \mathcal{F}_t(\chi) \) (2) for every \( \chi = (x^\top, \xi^\top, \theta^\top)^\top \in \mathbb{X} \),
\[
\tilde{c}_t(\chi) := -\min\{\text{trace}(\mathcal{F}_t(\chi)), b\}, \quad t \in \mathbb{T}_y, \quad (4)
\]
where \( b \in \mathbb{R}_+ \) is a large constant. One selects \( b \) empirically so that \( \min\{\text{trace}(\mathcal{F}_t(\chi)), b\} = \text{trace}(\mathcal{F}_t(\chi)) \) for all \( \chi \in \mathbb{X} \) of practical interest. \( b \) is needed so that \( \tilde{c}_t \) is b.b.

Using \( \tilde{c}_t \) (4) and \( \hat{c}_t \) (Sec. 2), we define stage cost functions \( c_t : \mathbb{X} \times \mathbb{C} \to \mathbb{R}^* \) for all \( t \in \mathbb{T} \) that include both information-theoretic and performance-based criteria:
\[
c_t(x, u) := \begin{cases} \hat{c}_t(x, u; \hat{\theta}) + \lambda \tilde{c}(\chi), & \text{if } t \in \mathbb{T}_y \setminus \{N\}, \\ c_t(x, u; \hat{\theta}), & \text{if } t \in \mathbb{T} \setminus \mathbb{T}_y, \end{cases} \quad (5a)
\]
\( \lambda \in \mathbb{R}_+ \) is chosen \textit{a priori} based on the relative importance of the two types of criteria. Similarly, we define a terminal cost function \( c_N : \mathbb{X} \to \mathbb{R}^* \) by
\[
c_N(\chi) := \begin{cases} \hat{c}_N(x; \hat{\theta}) + \lambda \tilde{c}(\chi), & \text{if } N \in \mathbb{T}_y, \\ \hat{c}_N(x; \hat{\theta}), & \text{if } N \notin \mathbb{T}_y. \end{cases} \quad (5b)
\]
A terminal cost is important for leukemia treatment because oncologists specify that the concentration of neutrophils should be within particular bounds by the end of a treatment cycle (to be exemplified in Sec. 6).

An advantage of incorporating the FIM into a cost function, as in (5), is that this choice avoids feasibility issues,
which may arise if the FIM is incorporated into a constraint. However, \( \lambda \) requires tuning. The authors of [11] propose a trade-off term so that their objective penalizes an expected loss of optimality, if the noise has sufficiently small bounded support [11, Th. 1]. In contrast, we permit unbounded noise. To formulate an optimal control problem using the cost functions (5), we will define the dynamics of \( \chi_t \) in the next subsection, starting with the dynamics of the parameter estimate \( \hat{\theta}_t \).

3.2 Defining \( \chi_t \)-dynamics

We define the dynamics of \( \hat{\theta}_t \) in every \( t \in T \) by

\[
\hat{\theta}_{t+1} = g_t(y_t, x_t; \hat{\theta}_t) := \begin{cases} 
\hat{g}_t(y_t, x_t; \hat{\theta}_t), & \text{if } t \in T_y \setminus \{N\}, \\
\hat{\theta}_t, & \text{if } t \in T \setminus T_y,
\end{cases}
\]

(6)

where \( \hat{\theta}_0 \) is an initial guess for \( \theta \). For every \( t \in T_y \setminus \{N\} \), we define \( \hat{g}_t \) using a gradient-based procedure [23, Eq. (5.73)] adapted to ensure that \( \hat{g}_t \) is \( \mathbb{P} \)-valued:

\[
\hat{g}_t(y, x; \theta) := \varphi_\epsilon(\theta - \alpha_t \cdot L_t(x; \theta)^{-1} \cdot D_\theta \hat{\epsilon}_t(x, y; \theta)).
\]

(7)

In (7), \( \varphi_\epsilon : \mathbb{R}^p \to \mathbb{P} \) is continuous (Sec. 1.2), \( \alpha_t \in \mathbb{R}_+ \) is a step size, \( D_\theta \hat{\epsilon}_t(x, y; \theta) \) is the gradient of the expectation of the squared measurement errors \( \hat{\epsilon}_t(x, y; \theta) := (y - h_t(x; \theta))^2 \) with respect to \( \theta \), and \( L_t(x; \theta) \in \mathbb{R}^{p \times p} \) is symmetric positive definite (to be specified). We may choose \( L_t(x; \theta) \) to be the identity matrix \( I_p \), so that (7) resembles gradient descent. Otherwise, we use a Gauss-Newton update with a regularization term \( \gamma \in \mathbb{R}_+ \) [23, Eq. (5.79)]:

\[
L_t(x; \theta) = 2(\epsilon)^\top \left( \frac{\partial h_t(x; \theta)}{\partial \theta} \right) + \gamma I_p.
\]

(8)

The \( \chi_t \)-dynamics depend on an output equation \( H_t \) and a state update equation \( F_t \). Using \( h_t \) (1) and an arbitrary vector \( v \in \mathbb{Y} \), we define \( H_t : \mathbb{X} \times \mathbb{Y} \to \mathbb{Y} \) by

\[
H_t(\chi, w) := \begin{cases} 
h_t(x; \hat{\theta}) + w, & \text{if } t \in T_y, \\
v, & \text{if } t \in T_N \setminus T_y.
\end{cases}
\]

(9)

\( v \) is a “dummy” vector that will be useful for writing the \( \chi_t \)-dynamics concisely and for analyzing these dynamics. We define \( F_t : \mathbb{X} \times \mathbb{C} \times \mathbb{D} \times \mathbb{Y} \to \mathbb{X} \) by

\[
F_t(\chi, u, d, w) := \begin{bmatrix} f_t(x, u, d; \hat{\theta}) \\ \phi_t(\xi, u, d; \hat{\theta}) \\ g_t(y, x; \hat{\theta}) \end{bmatrix}, \quad t \in T,
\]

(11)

where \( F_t \) depends on \( f_t \) (1), \( \phi_t \) (3), and \( g_t \) (6) as follows:

\[
F_t \begin{bmatrix} x \\ \xi \\ \hat{\theta} \end{bmatrix}, u, d, y := \begin{bmatrix} f_t(x, u, d; \hat{\theta}) \\ \phi_t(\xi, u, d; \hat{\theta}) \\ g_t(y, x; \hat{\theta}) \end{bmatrix}, \quad t \in T.
\]

Then, the \( \chi_t \)-dynamics are given by

\[
\chi_{t+1} = F_t(\chi_t, u_t, d_t, w_t), \quad t \in T,
\]

\[
y_t = H_t(\chi_t, w_t), \quad t \in T_N,
\]

(12)

where \( w_t \) is a realization of \( W_t \) if \( t \in T_y \), and \( w_t = v \) if \( t \in T_N \setminus T_y \). \( v \) does not affect \( \chi_{t+1} \) for any \( t \in T \). For every \( t \in T \setminus T_y \) and \((\chi, u, d, w) \in \mathbb{X} \times \mathbb{C} \times \mathbb{D} \times \mathbb{Y} \), we have that \( F_t(\chi, u, d, w) = G_t(\chi, u, d) \), where

\[
G_t(\chi, u, d) := \begin{bmatrix} f_t(x, u, d; \hat{\theta}) \\ \phi_t(\xi, x, u, d; \hat{\theta}) \\ \hat{\theta} \end{bmatrix}, \quad t \in T \setminus T_y.
\]

(13)

\( p \in \mathcal{P}_X \) denotes the prior distribution of the realizations \( \chi_0 \in \mathbb{X} \). \( p \) depends on the distribution of \( X_0 \) and the values of \( \xi_0 \) and \( \hat{\theta}_0 \) (see just below (3b) and (6), respectively). Let us now study the costs and dynamics presented above.

Theorem 1 (Regularity of \( c_t \), \( F_t \), and \( G_t \)) Let \( A1 \) hold. Then, \( c_t \) (5) is l.s.c. and b.b. for every \( t \in T_N \), and \( F_t \) (10) is continuous for every \( t \in T \). In particular, \( G_t \) (13) is continuous for every \( t \in T \setminus T_y \).

Proof. The first property holds because \( c_t \) is a sum of l.s.c. and b.b. functions, \( \hat{c}_t \) or \( \hat{c}_t + \lambda \hat{\epsilon}_t \), \( \lambda \hat{\epsilon}_t \) is l.s.c. because \( \hat{c}_t \) (4) is continuous under \( A1 \). \( \lambda \hat{\epsilon}_t \) is b.b. because \( \lambda \hat{\epsilon}_t(\chi) \geq -\lambda b > -\infty \) for every \( \chi \in \mathbb{X} \). For the second property, note that \( H_t \) (9) is continuous for every \( t \in T_N \) because it is either a sum of continuous functions or it is constant. Since \( F_t \) is a composition of \( F_t \) and \( H_t \), it suffices to show that \( F_t = (f_t, \phi_t, g_t) \) is continuous in each component. The continuity of \( f_t \) (1) and \( \phi_t \) (3b) follows from \( A1 \). There are two cases for \( g_t \) (6). If \( t \in T \setminus T_y \), then \( g_t(y, x; \hat{\theta}) = \hat{\theta} \), which is continuous. Otherwise, \( g_t = \hat{g}_t \) (7) depends on a continuous function \( \varphi_\epsilon \), a matrix inverse, and a gradient. The map \( (x, \hat{\theta}) \to L_t(x; \hat{\theta})^{-1} \) is continuous because \( L_t \) is defined by (8), \( L_t(x; \hat{\theta}) \) is positive definite for any \( (x, \hat{\theta}) \in \mathbb{S}_x \times \mathbb{P} \), and \( \frac{\partial h_t}{\partial \theta} \) is continuous under \( A1 \). The gradient \( D_\theta \hat{\epsilon}_t(y, x; \hat{\theta}) \in \mathbb{R}^p \) is given by

\[
D_\theta \hat{\epsilon}_t(y, x; \hat{\theta}) = -2 \left( \frac{\partial h_t(y, x; \hat{\theta})}{\partial \hat{\theta}} \right) (y - h_t(x; \hat{\theta})).
\]

(14)

The continuity of \( \frac{\partial h_t}{\partial \theta} \) and \( h_t \) under \( A1 \) imply the continuity of \( (y, x; \hat{\theta}) \to D_\theta \hat{\epsilon}_t(y, x; \hat{\theta}) \). Lastly, the continuity of \( \varphi_\epsilon \) and \( (y, x; \hat{\theta}) \to L_t(x; \hat{\theta})^{-1} \) imply the continuity of \( \hat{g}_t \) (7). \( \square \)

4 Representation as a regular POMDP

Here, we show that the system (12) equipped with the information-theoretic costs (5) is a nonstationary POMDP with regularity properties under \( A1 \). This representation will facilitate policy synthesis in Section 5. Next, we formalize the POMDP model of interest.

Definition 1 (Multi-time-scale regular POMDP)

A multi-time-scale regular POMDP consists of the following: (i) a state space \( \mathbb{X} \), compact control space \( \mathbb{C} \), and measurement space \( \mathbb{Y} \), which are nonempty Borel spaces; (ii) finite discrete-time horizons for the states and measurements, \( T_N \) and \( T_y \), respectively, satisfying
the definitions of Section 2; (iii) stage and terminal cost functions $c_t$ that are l.s.c. and b.b. for every $t \in T_N$; (iv) a distribution $p \in P_X$ of the initial state; (v) state transition kernels $q_t$ on $X$ given $X \times C$ that are weakly continuous for every $t \in \{1, 2, \ldots, N\}$; and (vi) observation kernels $s_t$ on $Y$ given $X$ that are continuous in total variation for every $t \in T_y$.

Under A1, the system (12) equipped with $c_t$ (5) satisfies Parts (i)-(iv) of Definition 1. We will define state transition and observation kernels and then show that they satisfy Parts (v)–(vi) of Definition 1 under A1. $\chi_t \in X$, $u_t \in C$, and $y_t \in Y$ are realizations of random objects $X_t$, $U_t$, and $Y_t$, respectively. For every $t \in T$, the state transition kernel $q_{t+1}$ provides the conditional distribution of $X_t$ given a realization of $(X_t, U_t)$.

**Definition 2 (State transition kernels $q_1, \ldots, q_N$)**

For every $t \in T_y \setminus \{N\}$, $q_{t+1}$ depends on the joint distribution $\nu_t$ of the process and measurement noise $(D_t, W_t)$ as follows: $q_{t+1}(\bar{X}, u) := \nu_t(\{(d, w) \in D \times Y : F_t(x, u, d, w) \in \bar{X}\})$, $t \in T_y \setminus \{N\}$,

for every $\bar{X} \in B_X$ and $(\chi, u) \in X \times C$. Otherwise, if $t \in T \setminus T_y$, then there is no measurement and thus no measurement noise. In this case, $q_{t+1}$ depends on the distribution $\mu_t$ of $D_t$ and the function $G_t$ (13):

$q_{t+1}(\bar{X}, u) := \mu_t(\{(d \in D : G_t(x, u, d) \in \bar{X})\})$, $t \in T \setminus T_y$,

for every $\bar{X} \in B_X$ and $(\chi, u) \in X \times C$.

$q_t$ differs from a typical state transition kernel in two ways. Its definition accounts for the absence of the measurement noise at particular time points and how this absence leads to dynamics functions with different dependencies. The observation kernel $s_t$ is simpler because it only pertains to $t \in T_y$. For every $t \in T_y$, $s_t$ provides the conditional distribution of $Y_t$ given a realization of $X_t$.

**Definition 3 (Observation kernel $s_t$ for $t \in T_y$)**

$s_t$ depends on the distribution $\rho_t$ of $W_t$ as follows:

$s_t(\bar{Y} | \chi) := \rho_t(\{w \in Y : h_t(x, \hat{\theta}) + w \in \bar{Y}\})$, $t \in T_y$,

for every $\bar{Y} \in B_Y$ and $\chi = (x^T, \xi^T, \hat{\theta}^T)^T \in X$.

In the POMDP literature, the observation kernel commonly depends on $u_t$. We do not allow $s_t$ to depend on $u_t$ to simplify the output sensitivity matrix (2b).

**Lemma 1 (Regularity of $q_t$ and $s_t$)**

Let A1 hold. Then, $q_t$ of Definition 2 and $s_t$ of Definition 3 satisfy Part (v) and Part (vi) of Definition 1, respectively.

**Proof.** For every $t \in T$, the weak continuity of $q_{t+1}$ follows from the continuity of $F_t$ (10) under A1 (Theorem 1). For every continuous bounded $\psi : X \to \mathbb{R}$, $(\chi, u) \in X \times C$, and $t \in T$, we have that $f_{\psi} \psi \ d\mu_{t+1}(|\chi, u) = \int_{\mathbb{R} \times \mathbb{R}} \psi(F_t(x, u, \cdot)) \ d\mu_t$, if $t \in T_y \setminus \{N\}$, \int_{\mathbb{R} \times \mathbb{R}} \psi(G_t(x, u, \cdot)) \ d\mu_t$, if $t \in T \setminus T_y$,

by a change-of-measure theorem [1, Th. 1.6.12]. Now, let $(\chi_n, u_n)_{n \in \mathbb{N}}$ in $X \times C$ converge to $(\chi, u) \in X \times C$, and let $\psi : X \to \mathbb{R}$ be continuous and bounded. It suffices to show that $\int_{\mathbb{R} \times \mathbb{R}} \psi \ d\mu_{t+1}(|\chi, u_n) \to \int_{\mathbb{R} \times \mathbb{R}} \psi \ d\mu_{t+1}(|\chi, u)$ in $\mathbb{R}$ as $n \to +\infty$ (Rmk. 1). First, consider $t \in T_y \setminus \{N\}$. Define the function $\psi_t : D \times Y \to \mathbb{R}$ by $\psi_{t,n} := \psi(F_t(x, u_n, \cdot))$ for every $n \in \mathbb{N}$, and define $\psi_t : D \times Y \to \mathbb{R}$ by $\psi_t := \psi(F_t(x, u, \cdot))$. We have that $\int_{D \times Y} \psi_t \ d\mu_t = \int_{D \times Y} \psi \ d\mu_{t+1}$ for every $n \in \mathbb{N}$ and $\int_{D \times Y} \psi_t \ d\mu_t = \int_{D \times Y} \psi \ d\mu_{t+1}$ for the first case of (15). Since $(\chi_n, u_n) \to (\chi, u)$ and $\psi \circ F_t$ is continuous, $\psi_{t,n} \to \psi_t$ pointwise as $n \to +\infty$. Also, $\psi$ is bounded, and $\psi_t$ and $\psi_{t,n}$ for every $n \in \mathbb{N}$ are continuous and hence measurable. Thus, we use the Dominated Convergence Theorem to conclude that $\int_{D \times Y} \psi_{t,n} \ d\mu_t = \int_{D \times Y} \psi \ d\mu_t$ in $\mathbb{R}$ as $n \to +\infty$ [1, p. 49]. The analysis of $q_{t+1}$ for any $t \in T := T \setminus T_y$ is similar, and so, we omit it. For every $t \in T_y$, the continuity of $s_t$ in total variation is due to $h_t$ being continuous, the measurement noise being additive, $\rho_t$ admitting a continuous nonnegative density, and Scheffé’s Lemma [35, Sec. 5.10]. The reader may see [21, Sec. 2.2 (ii)] or [5, Remark 5] for more details. \qed

Hence, the system (12) with the costs (5), state transition kernels of Definition 2, and observation kernels of Definition 3 is a POMDP that satisfies Definition 1 under A1.

### 5 Policy Synthesis

A POMDP can be reduced to a fully observable belief-space MDP, whose state space is the space of posterior distributions of the unobservable state [31] [36]. Moreover, if there exists an optimal policy for the belief-space MDP, then an optimal policy for the POMDP can be constructed [4, Prop. 10.3] [13, pp. 89–90]. Thus, the purpose of this section is to show the existence of an optimal policy under A1 for a belief-space MDP corresponding to the POMDP of the previous section. (Verifying the equivalence between the POMDP and belief-space MDP solutions is out of scope of the current brief paper; related proofs can be found in [4, Prop. 10.4, Prop. 10.5], for example.)

First, we formalize our belief-space MDP. We define its state and trajectory spaces (Def. 4), random states and controls (Def. 5), random cost variable (Def. 6), Markov policy class (Def. 7), state transition kernels (Def. 8, Def. 9), and initial distribution (Def. 10). Then, we provide a policy existence result (Theorem 2). The definitions are required to state the theorem. While the definitions resemble the standard ones for POMDPs [4, Ch. 10] [9] [13, Ch. 4], we must also circumvent the issue of measurements not necessarily being available always. This issue requires us to provide different definitions for
the state transition kernels (Def. 9) and for the initial distribution (Def. 10).

**Definition 4** $(Z, \Omega)$ The state space of the belief-space MDP is $Z := P_X$. A trajectory $\omega \in \Omega := (Z \times C)^N \times Z$ takes the form $\omega = (z_0, u_0, \ldots, z_{N-1}, u_{N-1}, z_N)$, where the coordinates of $\omega$ are related casually.

**Definition 5** $(Z_t, \bar{U}_t)$ We define $Z_t : \Omega \to Z$ for every $t \in T_N$ and $\bar{U}_t : \Omega \to C$ for every $t \in T$ to be projections, i.e., $Z_t(\omega) := z_t$ and $\bar{U}_t(\omega) := u_t$ for every $\omega \in \Omega$ of the form in Definition 4.

**Definition 6** $(C)$ Under A1, we define $C : \Omega \to \mathbb{R}^*$ by $C(\omega) := \tilde{c}_N(Z_N(\omega)) + \sum_{t=0}^{N-1} \tilde{c}_t(Z_t(\omega), \bar{U}_t(\omega))$, where $\tilde{c}_t$ is related to $c_t$ (5) as follows:

\[
\begin{align*}
\tilde{c}_t(z_t, u_t) &:= \int_X c_t(\chi_t, u_t) \, z_t(d\chi_t), \quad t \in T, \quad (16a) \\
\tilde{c}_N(z_N) &:= \int_X c_N(z_N, \chi_N) \, z_N(d\chi_N), \quad (16b)
\end{align*}
\]

with $\tilde{c}_t : Z \times C \to \mathbb{R}^*$ for every $t \in T$ and $\tilde{c}_N : Z \to \mathbb{R}^*$.

**Definition 7** $(\Pi, C_t)$ We define $C_t := C$ if $t \in T_u$ and $C_t := \{u\}$ if $t \in T \setminus T_u$. A control policy $\pi = (\pi_0, \pi_1, \ldots, \pi_{N-1}) \in \Pi$ is a finite sequence of measurable stochastic kernels on $C$ given $Z$ such that $\pi_t(C_t | z_t) = 1$ for every $z_t \in Z$ and $t \in T$. $\Pi$ is the set of all policies.

In Definitions 8 and 9, we specify the state transition kernels $\tilde{q}_t, \tilde{q}_2, \ldots, \tilde{q}_N$ of the belief-space MDP. For every $t \in T$, $\tilde{q}_{t+1}(z_t, u_t) \in P_Z$ is the conditional distribution of $Z_{t+1}$ given a realization $(z_t, u_t) \in Z \times C$ of $(Z_t, \bar{U}_t)$. Definition 8 applies when a measurement is present, whereas Definition 9 applies when a measurement is absent. In each case, some technical lead-up is required before writing the expression for $\tilde{q}_{t+1}$.

**Definition 8** ($\tilde{q}_{t+1}$ when a measurement is present) If $t + 1 \in \{1, 2, \ldots, N\} \setminus T_y$, then $\tilde{q}_{t+1}$ is constructed using a joint conditional distribution $q_{t+1}(\cdot | z_t, u_t)$ of $(Y_{t+1}, \chi_{t+1})$. $q_{t+1}$ is a measurable stochastic kernel on $Z \times X$ given $Z \times \tilde{C}$ such that $q_{t+1}(\widetilde{X} | \tilde{z}_t, u_t) :=$ 

\[
\int_X \int_{\tilde{X}} \text{r}_{t+1}(\chi_{t+1} | \tilde{z}_t, u_t) \, \text{q}_{t+1}(d\chi_{t+1} | x_{t+1}) \, z_t(d\chi_t) \quad (17)
\]

for every $X, \tilde{X} \in B_X$, $\tilde{X} \in B_{\tilde{X}}$, and $(z_t, u_t) \in Z \times C$. Denote the marginal of $r_{t+1}$ on $Y$ by $r_{t+1}'$, i.e., $r_{t+1}'(\tilde{X} | z_t, u_t) := r_{t+1}(X \times \tilde{X} | z_t, u_t)$ for every $X \in B_X$ and $(z_t, u_t) \in Z \times C$. $r_{t+1}$ enjoys a decomposition in terms of this marginal as follows: for every $Y, X, \tilde{X} \in B_Y$, $X \in B_X$, and $(z_t, u_t) \in Z \times C$, $r_{t+1}(X \times \tilde{X} | z_t, u_t) = \int_{\tilde{X}} \text{q}_{t+1}(\tilde{X} | z_t, u_t, y) \, \text{r}_{t+1}'(dy | z_t, u_t)$, where $\text{q}_{t+1}$ is a measurable stochastic kernel on $X$ given $Z \times \tilde{C}$ [4, Cor. 7.27.1]. Then, we define $\tilde{q}_{t+1}(\tilde{X} | z_t, u_t) :=$ 

\[
r_{t+1}'(\{y_{t+1} \in Y : \text{q}_{t+1}(\cdot | z_t, u_t, y_{t+1}) \in \tilde{X}\} | z_t, u_t) \quad (18)
\]

for every $Z, \tilde{X} \in B_Z$ and $(z_t, u_t) \in Z \times C$.

**Definition 9** ($\hat{q}_{t+1}$ when a measurement is absent) If $t + 1 \in \{1, 2, \ldots, N\} \setminus T_y$, then $\hat{q}_{t+1}$ concentrates the realizations of $Z_{t+1}$ at a prior conditional distribution $\eta_{t+1}(\cdot | z_t, u_t) \in Z$ of $X_{t+1}$, i.e., we define $\hat{q}_{t+1}$ by

\[
\hat{q}_{t+1}(\tilde{X} | z_t, u_t) := \delta_{\eta_{t+1}(\cdot | z_t, u_t)}(\tilde{X}) \quad (19)
\]

for every $\tilde{X} \in B_Z$ and $(z_t, u_t) \in Z \times C$, where $\eta_{t+1}$ is a measurable stochastic kernel on $X$ given $Z \times C$, which is defined by [4, p. 261, Eq. (50)]

\[
\eta_{t+1}(\tilde{X} | z_t, u_t) := \int_X q_{t+1}(X | \chi_t, u_t) \, z_t(d\chi_t) \quad (20)
\]

for every $\tilde{X} \in B_{\tilde{X}}$ and $(z_t, u_t) \in Z \times C$.

For $t+1 \in \{1, 2, \ldots, N\}$, the expression for $\hat{q}_{t+1}$ is in (18) or (19). Lastly, we specify the initial kernel $\hat{q}_0$, whose construction also depends on the availability of a measurement. $\hat{q}_0(\cdot | p) \in P_Z$ is the distribution of $Z_0$ when $p \in Z$ is the distribution of $X_0$.

**Definition 10** ($\hat{q}_0$) If $0 \in T_y$, then $\hat{q}_0$ depends on a joint distribution $r_0(\cdot, \cdot | p)$ of $(Y_0, X_0)$, where $r_0$ is defined by

\[
r_0(Y \times X | p) := \int_X s_0(\tilde{X} | x, p) \, r_0(dy | p) \quad (21)
\]

\[
= \int_X \phi_0(X | p, y_0) \, r_0(dy | p) \quad (22)
\]

for every $Y \in B_Y$, $X \in B_X$, and $p \in Z$, $r_0$ is defined by

\[
r_0(Y | p) := r_0(Y \times X | p) \quad (23)
\]

for every $Y \in B_Y$, $X \in B_X$, and $p \in Z$.

Equipped with the above definitions, now we can specify a probability measure, an expected cost, and the optimal control problem of interest. By the Ionescu-Tulcea Theorem, for every $\pi \in \Pi$ and $p \in Z$, there is a unique probability measure $P_{\phi_0}(\pi | p) \in P_Z$ that depends on $\pi$, $\hat{q}_0(\cdot | p)$, and $\hat{q}_{t+1}$ for all $t \in T$. If $G : \Omega \to \mathbb{R}^*$ is measurable and b.b., then the expectation of $G$ with respect to $P_{\phi_0}(\pi | p)$ is defined by $E_{\phi_0}(\pi | p)(G) := \int_G G \, dP_{\phi_0}(\pi | p)$. The next result guarantees that an optimal policy $\pi^* \in \Pi$ for the belief-space MDP exists under A1, i.e.,

\[
E_{\phi_0}(\pi^* | p)(C) = J^*(p) := \inf_{\pi \in \Pi} E_{\phi_0}(\pi | p)(C), \quad p \in Z. \quad (24)
\]

**Theorem 2** (Policy synthesis) Let A1 hold. Define $J_t := \tilde{c}_N$ (16b), and for $t = N - 1, \ldots, 0$, define $J_t : Z \to \mathbb{R}^*$ recursively backwards in time by

\[
J_t(z_t) := \inf_{u_t \in C_t} V_t(z_t, u_t), \quad (25)
\]

where $C_t$ is from Def. 7, $V_t := \tilde{c}_t + V_{t+1}'$, $\tilde{c}_t$ is defined by (16a), and $V_{t+1}' : Z \times C \to \mathbb{R}^*$ depends on $J_{t+1}$ as follows:
\[ V'_t(z_t, u_t) := \int_{\mathcal{Z}} J_{t+1}(z_{t+1}) \; \tilde{q}_{t+1}(dz_{t+1}|z_t, u_t). \quad (26) \]

Then, for every \( t \in T_N \), \( J_t \) is l.s.c. and b.b. Also, for every \( t \in T \), there is a measurable function \( \kappa^*_t : \mathcal{Z} \rightarrow \mathbb{C} \) such that \( \kappa^*_t(z) \in \mathcal{C}_z \) and \( J_t(z_t) = V_t(z_t, \kappa^*_t(z_t)) \) for every \( z_t \in \mathcal{Z} \). In particular, for every \( t \in T \setminus T_n \), we choose \( \kappa^*_t(z_t) := u \) for every \( z_t \in \mathcal{Z} \). Lastly, \( \pi^* := (\delta^{*\tau}, \delta^{*\eta}, \ldots, \delta^{*\tau}) \in \Pi \) (Rmk. 4) is optimal for the belief-space MDP, i.e., \( \pi^* \) satisfies (24).

**Proof.** We will prove the first two statements by induction. The arguments have nuances due to the two cases for \( \tilde{q}_{t+1} \) (18) (19). The last statement holds by a typical dynamic programming argument, e.g., see [4, Ch. 8] or [14, Ch. 3].

Under A1, for every \( t \in T_N \), \( \tilde{c} \) (16) is l.s.c. and b.b. because \( c_t \) (5) enjoys these properties (Theorem 1) and due to the Generalized Fatou’s Lemma [9, Lemma 6.1]. Now, \( J_N = \tilde{c} \) is l.s.c. and b.b. Assume the induction hypothesis: for some \( t \in T \), \( J_{t+1} \) is l.s.c. and b.b. If \( V'_t \) (26) is l.s.c. and b.b., then \( J_t \) (25) is l.s.c. and b.b., and there is a measurable function \( \kappa^*_t \) that satisfies the properties specified in the statement of Theorem 2 (A1, Rmk. 5). If \( \tilde{q}_{t+1} \) is weakly continuous, then \( V'_t \) (26) is l.s.c. and b.b. (induction hypothesis, Rmk. 3).

Hence, it suffices to show that \( \tilde{q}_{t+1} \) is weakly continuous. First, consider \( t + 1 \in T_N \setminus \{0\} \). Then, \( q_{t+1} \) (Def. 2) is weakly continuous, and \( \tilde{q}_{t+1} \) (Def. 3) is continuous in total variation under A1 (Lemma 1). These two continuity properties imply that \( \tilde{q}_{t+1} \) (18) is weakly continuous directly from [21, Th. 1], which first appeared in [9]. Otherwise, if \( t + 1 \in \{1, 2, \ldots, N\} \setminus T_N \), then \( \tilde{q}_{t+1} \) is defined by (19). We must show that \((z, u) \rightarrow \tilde{q}_{t+1}(\cdot|z, u) = \delta_{\eta_{t+1}}(\cdot|z, u) \) is weakly continuous. We define \( \ell : \mathcal{Z} \rightarrow \mathcal{P}_2 \) by \( \ell(z) := \delta_z \) and \( v : \mathcal{Z} \times \mathcal{C} \rightarrow \mathcal{Z} \) by \( v(z, u) := \eta_{t+1}(z, u) \), and thus, \( \tilde{q}_{t+1}(\cdot|z, u) = \ell(v(z, u)) \). Also, \( \ell \) is weakly continuous (Rmk. 3). Hence, it suffices to show that \( v \) is weakly continuous. More specifically, it suffices to show that for any continuous bounded function \( \psi : \mathcal{X} \rightarrow \mathbb{R} \), the function \( \psi : \mathcal{Z} \times \mathcal{C} \rightarrow \mathbb{R} \) defined by \( \psi(z, u) := \int_{\mathcal{X}} \psi d\eta_{t+1}(\cdot|z, u) \) is continuous (Rmk. 1).

By the definition of \( \eta_{t+1} \) (20), we have that
\[ \psi(z, u) = \int_{\mathcal{X}} \int_{\mathcal{C}} \psi(a) q_{t+1}(da|x, u) \; z(dx). \quad (27) \]

We will re-express \( \psi' \) (27) to apply the weak continuity of products of measures in \( \mathcal{P}_X \) and \( \mathcal{P}_C \). Define \( \Psi : \mathcal{X} \times \mathcal{C} \times \mathcal{X} \rightarrow \mathbb{R} \) by \( \Psi(x, u, \cdot) := \psi \) for all \((x, u) \in \mathcal{X} \times \mathcal{C} \). Next, define \( \vartheta : \mathcal{X} \times \mathcal{C} \rightarrow \mathcal{X} \) by
\[ \vartheta(x, u) := \int_{\mathcal{X}} \Psi(x, u, a) \; q_{t+1}(da|x, u), \quad (28) \]
which is continuous and bounded (Lemma 1, Rmk. 3).

Now, \( \psi' \) (27) can be expressed in terms of \( \vartheta \) (28):
\[ \psi'(z, u) = \int_{\mathcal{X}} \int_{\mathcal{C}} \vartheta(x, u) \; \vartheta_{t+1}(dx) \; z(dx) \quad (29) \]
\[ = \int_{\mathcal{X} \times \mathcal{C}} \vartheta(dz \vartheta_{t+1}), \quad (30) \]
where \( z \vartheta_{t+1} \in \mathcal{P}_{X \times C} \) is the product of \( z \in \mathcal{P}_X \) and \( \vartheta_{t+1} \in \mathcal{P}_C \). Let \( z_n \rightarrow z \) weakly in \( \mathcal{Z} \), and let \( u_n \rightarrow u \) in \( \mathcal{C} \). The latter implies that \( \vartheta_{t+1} \rightarrow \vartheta_{t+1} \) weakly in \( \mathcal{P}_C \) (Rmk. 3). The weak convergence of \( z_n \rightarrow z \) in \( \mathcal{X} \) and \( \vartheta_{t+1} \rightarrow \vartheta_{t+1} \) in \( \mathcal{P}_C \) implies the weak convergence of \( z_n \vartheta_{t+1} \rightarrow z \vartheta_{t+1} \) in \( \mathcal{P}_{X \times C} \) (Rmk. 3).

The weak convergence of these products is equivalent to the convergence of \( \int_{\mathcal{X} \times \mathcal{C}} h d(z_n \vartheta_{t+1}) \rightarrow \int_{\mathcal{X} \times \mathcal{C}} h d(z \vartheta_{t+1}) \) in \( \mathbb{R} \) for every continuous bounded function \( h : \mathcal{X} \times \mathcal{C} \rightarrow \mathbb{R} \) (Rmk. 1). By considering \( h = \vartheta \) (28), we find that \( \psi'(z_n, u_n) \rightarrow \psi'(z, u) \in \mathbb{R} \) in view of (30). □

This concludes the theoretical portion of the brief paper.

### 6 Conceptual example

In this section, we provide a conceptual leukemia treatment model and explain why it satisfies A1. We consider the setting of adjusting the dose of an oral chemotherapy called 6-Mercaptopurine (6-MP) during a treatment cycle for a leukemia patient. Currently, the dose is adjusted according to the current neutrophil concentration (a neutrophil is a type of white blood cell). The concentration is measured at least once per cycle, whose typical length is 21 days. If the concentration is outside a desired range, then the dose of 6-MP is modified (increased or decreased by 20%) for the next cycle. Namely, the drug is taken at the prescribed dose for the first 14 days, and then no drug is taken for the remaining 7 days.

Continuous-time models for the kinetics of 6-MP [17] [18] and their effect on white blood cells [18] [22] have been developed recently. The models are deterministic in [17] [18]. Three states represent the amounts of 6-MP-derived substances in [17]. Five states represent the amounts of white blood cells in different stages of maturity under the influence of 6-MP in [18]. A stochastic model with two states that neglects the delayed effect of 6-MP on white blood cells has been proposed in [22]. These models represent one or both of the following biochemical processes: 1) how 6-MP is broken down into a substance called 6-thioguanine nucleotide (6-TGN); 2) how 6-TGN influences the proliferation and maturation of white blood cells in the bone marrow.
are the states representing the amounts of 6-MP-derived substances from [17], while \( x_4 \) through \( x_8 \) are the five states representing the amounts of white blood cells in different stages of maturity under the influence of 6-MP from [18] (Table 1). The parameters include various rates, drug-effect quantities, and the ratio between neutrophils and mature white blood cells (Table 1). The measurements are neutrophils and mature white blood cells. The spaces of interest are \( \mathbb{S} = \mathbb{R}_+^8, \mathbb{P} = \mathbb{R}^{13}, \mathbb{C} = [0, \bar{u}], \mathbb{D} = \mathbb{R}_+, \) and \( \mathbb{Y} = \mathbb{R}^2 \).

| Symbol | Description |
|--------|-------------|
| \( x_1 \) | Amount of 6-MP in the gut |
| \( x_2 \) | Amount of 6-MP in the plasma |
| \( x_3 \) | Concentration of 6-TGN in red blood cells |
| \( x_4 \) | Concentration of proliferating white blood cells |
| \( x_5, x_6, x_7 \) | Concentrations of maturing white blood cells |
| \( x_8 \) | Concentration of mature white blood cells |
| \( \theta_1 \) | Conversion rate (6-MP to 6-TGN) |
| \( \theta_2 \) | Michaelis-Menten constant |
| \( \theta_3 \) | Maximum proliferation rate |
| \( \theta_4 \) | Steepness parameter |
| \( \theta_5 \) | Feedback parameter |
| \( \theta_6 \) | Maximum drug effect on mature white blood cells |
| \( \theta_7 \) | Saturation constant for drug effect |
| \( \theta_8, \ldots, \theta_{12} \) | Rate of absorption, elimination, or transition |
| \( \theta_{13} \) | Ratio between neutrophils and mature white blood cells |

Table 1

Letting \( \Delta \in \mathbb{R}_+ \) be the duration of \([t, t + 1]\) for every \( t \in \mathbb{T} \) and \( \hat{\varphi}_e : \mathbb{R}^8 \to \mathbb{S} \) be a continuously differentiable approximation for \( \varphi_e : \mathbb{R}^8 \to \mathbb{S} \) (Sec. 1.2), we consider

\[
\hat{f}_i(x, u, d; \theta) = \hat{\varphi}_e(x + \Delta \cdot \hat{f}(x, u; \theta) + d),
\]

\[
h_i(x; \theta) = C(\theta)x,
\]

\[
\hat{f}(x, u; \theta) := A(\theta)x + Bu + \hat{f}(x; \theta),
\]

where \( A(\theta) \in \mathbb{R}^{8 \times 8} \) and \( C(\theta) \in \mathbb{R}^{2 \times 8} \) depend linearly on \( \theta \), \( B \in \mathbb{R}^8 \) is constant, and \( \hat{f} \) is nonlinear. The definition of \( \hat{f} \) (33) comes from [17] and [18]. The nonlinear part is

\[
\hat{f} := [0, \hat{f}_2, -\hat{f}_2, \hat{f}_4, 0, 0, 0, 0]^\top,
\]

\[
\hat{f}_2(x; \theta) := -\theta_1 x_2 (x_2 + x_5)^{-1},
\]

\[
\hat{f}_4(x; \theta) := \left( \frac{\theta_3}{1 + (x_8/\theta_3)^{\theta_4}} - \frac{\theta_6 x_3}{\theta_7 + x_3} \right) x_4.
\]

\( \hat{f}_2 \) describes how 6-MP in the plasma \( (x_2) \) is broken down into 6-TGN \( (x_3) \). \( \hat{f}_4 \) describes how the expansion rate of the proliferating white blood cells \( (x_4) \) is influenced by the mature white blood cells \( (x_8) \) and the drug-derived substance \( x_3 \). The linear part \( A(\theta)x + Bu \) is given by

\[
A(\theta)x + Bu = \begin{bmatrix}
-\theta_6 x_1 + u \\
\theta_6 x_1 - \theta_6 x_2 \\
-\theta_{10} x_3 \\
-\theta_{11} x_4 \\
\theta_{11} (x_4 - x_5) \\
\theta_{11} (x_5 - x_6) \\
\theta_{11} (x_6 - x_7) \\
\theta_{11} x_7 - \theta_{12} x_8
\end{bmatrix},
\]

which includes the ingestion of 6-MP and the proliferation and maturation of the white blood cells. One can show that \( f_1 \) (31) satisfies Parts (a) and (f) of A1. The amounts of mature white blood cells \( (x_8) \) and neutrophils \( (\theta_{13} x_8) \) are relevant for clinical decisions. We choose \( y \in \mathbb{R}^2 \), where \( y_1 \) and \( y_2 \) are the measured amounts of mature white blood cells and neutrophils, respectively, and thus,

\[
C(\theta) = \begin{bmatrix}
0 & \cdots & 0 & 1 \\
0 & \cdots & 0 & \theta_{13}
\end{bmatrix} \in \mathbb{R}^{2 \times 8},
\]

where \( \theta_{13} \in (0, 1) \) is the ratio between neutrophils and mature white blood cells (Table 1). The measurement equation \( h_t \) (32) satisfies Parts (d) and (f) of A1.

Lastly, we specify cost functions \( \hat{c}_t \) to penalize (i) deviations of the amount of neutrophils with respect to a desired range \( [\bar{b}, \bar{b}] \) and (ii) low doses of 6-MP. The first
criterion serves to protect the patient’s immune system, while the second criterion serves to limit the production of cancer cells. For example, we may choose
\begin{align}
\hat{c}_t(x, u; \theta) &:= \frac{1}{N_\theta} (\zeta(x; \theta) - \lambda u^2), \quad t \in T, \\
\hat{c}_N(x; \theta) &:= \frac{1}{N_\theta} \zeta(x; \theta), \\
\zeta(x; \theta) &:= (\theta_{13} x_8 - \hat{b}) (\theta_{13} x_8 - \bar{b}),
\end{align}
with \( \lambda \in \mathbb{R}_+ \), where \( \zeta \) penalizes deviations of the amount of neutrophils with respect to \([\bar{b} \ \hat{b}] \), \( \hat{c}_t \) (38) is continuous and hence L.s.c. for every \( t \in T_N \). \( \zeta \) is b.b. since it is convex quadratic. Since \( 0 \leq u \leq \bar{u} \), we have that \(-\lambda u^2 \geq -\lambda \bar{u}^2 \). Thus, \( \hat{c}_0, \hat{c}_1, \ldots, \hat{c}_N \) (38) satisfy Part (b) of A1.

7 Conclusion
We have provided a theoretical foundation that unifies several features of biomedical systems, e.g., multiple time scales, partial observability, and uncertain parameters. A future theoretical step is to analyze the performance of the control policy from Theorem 2 under the true dynamics (1), which depend on \( \theta \) true. A shortcoming of Section 6 is that the partial observability and high dimensionality of the model (which is based on models from [17] and [18]) preclude the practical application of Theorem 2. Indeed, POMDPs are notorious for their numerical complexity, except in special cases such as the linear-quadratic-Gaussian case, which does not apply here. Hence, one aim of our ongoing work is the development of methodology with reduced complexity. We are collecting leukemia patient data, which can be useful for reducing the parameter space (i.e., use the data to identify the subset of parameters that are common to the patient population, as in, e.g., [20]). Taking inspiration from the framework proposed here, we hope to devise new methods that are resilient to process and measurement noise but also are computationally tractable, which may be facilitated by patient data and domain-specific modeling structure.

In the long term, we envision a technology that provides optimized doses for the next cycle given the patient’s prior doses and measurements. During a clinical visit, the patient and her oncologist would see on a computer screen how the predicted trajectories of mature white blood cells and neutrophils refit to the patient’s data set when a new measurement becomes available. The optimized doses would be compared to the typical doses visually, and then the oncologist would decide which doses to prescribe for the next cycle. The technology may benefit from the theoretical foundation that we have developed here to conceptually unify the unobservable state, unknown parameters, multiple time scales, and nonlinear stochastic dynamics.

Acknowledgements
M.P.C. appreciates Chunming Wei, Zhengmin Yang, Zehua Li, and Huizhen Yu for fruitful discussions, and she appreciates support provided by the Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto. The authors wish to thank two anonymous reviewers for their helpful suggestions.

References
[1] R. B. Ash. Real Analysis and Probability. Academic Press, Inc., New York, NY, USA, 1972.
[2] K. J. Åström and B. Wittenmark. Adaptive Control, 2nd edition. Dover Publications, Inc., Mineola, NY, USA, 1995.
[3] V. A. Bavdekar, V. Ehlinger, D. Gidon, and A. Mesbah. Stochastic predictive control with adaptive model maintenance. In Proc. IEEE Conf. Decis. Control, pages 2745–2750. IEEE, 2016.
[4] D. P. Bertsekas and S. E. Shreve. Stochastic Optimal Control: The Discrete-Time Case. Athena Scientific, Belmont, MA, USA, 1996.
[5] M. P. Chapman, R. Bonalli, K. M. Smith, I. Yang, M. Pavone, and C. J. Tomlin. Risk-sensitive safety analysis using conditional value-at-risk. IEEE Trans. Autom. Control, 67(12):6521–6536, 2022.
[6] D. J. DeAngelo, K. E. Stevenson, S. E. Dahlberg, L. B. Silverman, S. Couban, J. G. Supko, et al. Long-term outcome of a pediatric-inspired regimen used for adults aged 18–50 years with newly diagnosed acute lymphoblastic leukemia. Leukemia, 29(3):526–534, 2015.
[7] N. Eling, M. D. Morgan, and J. C. Marioni. Challenges in measuring and understanding biological noise. Nat. Rev. Genet., 20(9):536–548, 2019.
[8] M. K. Erdal, K. W. Plaxco, J. Gerson, T. E. Kippin, and J. P. Hespanha. Optimal experiment design with applications to pharmacokinetic modeling. In Proc. IEEE Conf. Decis. Control, pages 3072–3079. IEEE, 2021.
[9] E. A. Feinberg, P. O. Kasyanov, and M. Z. Zgurovsky. Partially observable total-cost Markov decision processes with weakly continuous transition probabilities. Math. Oper. Res., 41(2):656–681, 2016.
[10] E. A. Feinberg, P. O. Kasyanov, and M. Z. Zgurovsky. Markov decision processes with incomplete information and semiuniform Feller transition probabilities. SIAM J. Control Optim., 60(4):2488–2513, 2022.
[11] X. Feng and B. Hou. Real-time algorithm for self-reflective model predictive control. J. Process Control, 65:68–77, 2018.
[12] T. A. N. Heirung, B. E. Ydstie, and B. Foss. Dual adaptive model predictive control. Automatica, 80:340–348, 2017.
[13] O. Hernández-Lerma. Adaptive Markov Control Processes. Springer-Verlag, New York, NY, USA, 1989.
[14] O. Hernández-Lerma and J. B. Lasserre. Discrete-Time Markov Control Processes: Basic Optimality Criteria. Springer-Verlag, New York, NY, USA, 1996.
[15] L. Hewing, K. P. Wabersich, M. Menner, and M. N. Zeilinger. Learning-based model predictive control: Toward safe learning in control. Annu. Rev. Control Robot. Auton. Syst., 3:269–296, 2020.
[16] K. Hinderer. Foundations of Non-stationary Dynamic Programming with Discrete Time Parameter. Springer-Verlag, Berlin, Germany, 1970.
[17] D. Jayachandran, J. Lainez-Aguirre, A. Randell, T. Vik, R. Hannemann, G. Reklaitis, and D. Ramkrishna. Model-based individualized treatment of chemotherapeutics: Bayesian population modeling and dose optimization. PLoS One, 10(7):e0133244, 2015.
[18] D. Jayachandran, A. E. Rundell, R. E. Hannemann, T. A. Vik, and D. Ramkrishna. Optimal chemotherapy for leukemia: A model-based strategy for individualized treatment. *PLoS One*, 9(10):e109623, 2014.

[19] F. Jost. *Model-based optimal treatment schedules for acute leukemia*. PhD thesis, Otto-von-Guericke-Universität Magdeburg, Magdeburg, Germany, 2020.

[20] F. Jost, J. Zierk, T. T. Le, T. Raupach, M. Rauh, M. Sutterp, M. Stanulla, M. Metzler, and S. Sager. Model-based simulation of maintenance therapy of childhood acute lymphoblastic leukemia. *Front. Physiol.*, 11:217, 2020.

[21] A. D. Kara, N. Saldi, and S. Yüksel. Weak Feller property of non-linear filters. *Syst. Control. Lett.*, 134:104512, 2019.

[22] S. Karppinen, O. Lohi, and M. Vihola. Prediction of leukocyte counts during paediatric acute lymphoblastic leukaemia maintenance therapy. *Sci. Rep.*, 9(1):1–11, 2019.

[23] K. J. Keesman. *System Identification: An Introduction*. Springer-Verlag, London, U.K., 2011.

[24] H. C. La, A. Potschka, J. P. Schlöder, and H. G. Bock. Dual control and online optimal experimental design. *SIAM J. Sci. Comput.*, 39(4):B640–B657, 2017.

[25] C. A. Larsson, A. Elbadawi, C. R. Rojas, X. Bombois, and H. Hjalmarsson. An application-oriented approach to dual control with excitation for closed-loop identification. *Eur. J. Control*, 29:1–16, 2016.

[26] L. Ljung. *System Identification: Theory for the User*. Prentice Hall, Hoboken, NJ, USA, 1999.

[27] A. Mesbah. Stochastic model predictive control with active uncertainty learning: A survey on dual control. *Annu. Rev. Control*, 45:107–117, 2018.

[28] P. Nimmegeers, S. Bhonsale, D. Telen, and J. van Impe. Optimal experiment design under parametric uncertainty: A comparison of a sensitivities based approach versus a polynomial chaos based stochastic approach. *Chem. Eng. Sci.*, 221:115651, 2020.

[29] L. Pronzato. Optimal experimental design and some related control problems. *Automatica*, 44(2):303–325, 2008.

[30] C. V. Rao, D. M. Wolf, and A. P. Arkin. Control, exploitation and tolerance of intracellular noise. *Nature*, 420(6912):231–237, 2002.

[31] D. Rhenius. Incomplete information in Markovian decision models. *Ann. Stat.*, pages 1327–1334, 1974.

[32] S. Sager. Sampling decisions in optimum experimental design in the light of Pontryagin’s maximum principle. *SIAM J. Control Optim.*, 51(4):3181–3207, 2013.

[33] D. Telen, B. Houska, F. Logist, E. van Derlinden, M. Diehl, and J. van Impe. Optimal experiment design under process noise using Riccati differential equations. *J. Process Control*, 23(4):613–629, 2013.

[34] D. Telen, B. Houska, M. Vallerio, F. Logist, and J. van Impe. A study of integrated experiment design for NMPC applied to the Droop model. *Chem. Eng. Sci.*, 160:370–383, 2017.

[35] D. Williams. *Probability with Martingales*. Cambridge University Press, Cambridge, U.K., 1991.

[36] A. A. Yushkevich. Reduction of a controlled Markov model with incomplete data to a problem with complete information in the case of Borel state and control spaces. *Theory Probab. Its Appl.*, 21(1):153–158, 1976.