Sub-optimality bounds for certainty equivalent policies in partially observed systems

Berk Bozkurt, Aditya Mahajan, Ashutosh Nayyar, and Yi Ouyang

***Abstract*—In this paper, we present a generalization of the certainty equivalence principle of stochastic control. One interpretation of the classical certainty equivalence principle for linear systems with output feedback and quadratic costs is as follows: the optimal action at each time is obtained by evaluating the optimal state-feedback policy of the stochastic linear system at the minimum mean square error (MMSE) estimate of the state. Motivated by this interpretation, we consider certainty equivalent policies for general (non-linear) partially observed stochastic systems that allow for any state estimate rather than restricting to MMSE estimates. In such settings, the certainty equivalent policy is not optimal. For models where the cost and the dynamics are smooth in an appropriate sense, we derive upper bounds on the sub-optimality of certainty equivalent policies. We present several examples to illustrate the results.**

I. Introduction

In many applications in robotics, autonomous systems, finance, healthcare, and other domains the decision maker does not have access to the complete state of the system. Such systems are often modeled as partially observable Markov decision processes (POMDPs). The standard approach for solving POMDPs is to translate them into fully observed Markov decision processes (MDPs) by considering the posterior belief of the decision maker on the current state as a sufficient statistic [2], [3]. There is a rich literature on algorithms which use the structure of the resulting belief-state MDP to obtain optimal and approximately optimal policies.

However, finding the optimal policy is PSPACE-hard [4]. Most algorithms to find optimal policies have exponential worst-case complexity in the size of the state and observation spaces making them impractical for large-scale problems. Meanwhile, many heuristic approaches such as point-based value iteration methods [5] can be computationally efficient but lack provable performance guarantees. In fact, finding approximately optimal policies is also PSPACE-hard [6], [7], indicating that general purpose algorithms may not be efficient for all POMDP models.

These challenges have motivated significant interest in identifying structured classes of policies that are both computationally tractable and have good performance guarantees. Examples include policies based on a finite window of previous observations [8] (called frame stacking in reinforcement learning) and, more generally, policies based on *agent state*, which is a recursively updatable function of the past observations and actions [9], [10]. Several papers have identified sufficient conditions for such policies to perform well, including approximate information state [11], filter stability [12]–[14], weakly revealing observations [15], low covering numbers [16], low-rank structure [17], and revealing observation models [18]. These results highlight that structured policies can be approximately optimal for a specific sub-class of POMDPs.

In linear systems with quadratic cost and Gaussian noise (the so-called LQG problem), the optimal policy may be viewed as a structured policy with the following structure: the optimal action at each time is a linear function of the MMSE (minimum mean square error) state estimate and the corresponding feedback gain is the same as the feedback gain of the optimal *state-feedback* control of the *deterministic* system obtained by replacing all random variables by their means. This result is typically called the *certainty equivalence principle of stochastic control* [19]–[21] and has been generalized to other settings including systems with non-linear dynamics [22]–[24] and risk-sensitive control [25].

In this paper, we present a generalization of the certainty equivalence principle to general POMDPs. Our generalization is based on a slightly different interpretation of the LQG certainty equivalence principle: consider the optimal policy of the *stochastic* system with perfectly observable states and evaluate that policy at the MMSE state estimate. Similar views on the certainty equivalence principle have been used in the reinforcement learning and adaptive control literature [26] and are sometimes called partially stochastic certainty equivalence [27]. For clarity, we present a formal description of this interpretation.

Let $\mathcal{P}$ denote the partially observable linear system with state $s_t \in \mathcal{S}$, action $a_t \in \mathcal{A}$, and output $y_t \in \mathcal{Y}$, where $\mathcal{S}$, $\mathcal{A}$, and $\mathcal{Y}$ are Euclidean spaces. Let $\mathcal{M}$ be the fully observable stochastic linear system where the decision maker has access to the state. Note that the fully observed system $\mathcal{M}$ is different from one typically assumed in certainty equivalence. As is the case in the standard certainty equivalence principle, we are assuming that $\mathcal{M}$ is fully observed but we are not assuming that the dynamics of $\mathcal{M}$ are deterministic. For simplicity, suppose that the system runs for a finite horizon T. Let $\pi^{\mathcal{M}} = (\pi_1^{\mathcal{M}}, \dots, \pi_T^{\mathcal{M}})$ denote the optimal policy for model $\mathcal{M}$ and $\mu^{\mathcal{P}} = (\mu_1^{\mathcal{P}}, \dots, \mu_T^{\mathcal{P}})$ denote the optimal policy for model $\mathcal{P}$. Moreover, for any history $h_t = (y_1, a_1, y_2, a_2, \dots, y_t)$ of

A preliminary version of this work appeared in [1].

Berk Bozkurt is with INLAN, Montreal, QC, Canada. (email: berk.bozkurt@mail.mcgill.ca).

Aditya Mahajan is with the Department of Electrical and Computer Engineering, McGill University, Montreal, QC, Canada. (email: aditya.mahajan@mcgill.ca).

Ashutosh Nayyar is with the Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA, USA. (email: ashutosn@usc.edu).

Yi Ouyang is with Atmanity, Santa Clara, CA, USA. (email: ouyangyii@gmail.com).

observations and actions until time t, let $\mathcal{E}_t(h_t)$ denote the MMSE estimate of the state given the history h_t. Then, the standard result for LQG optimal control is that

$$\mu^{\mathcal{P}}_t(h_t) = \pi^{\mathcal{M}}_t(\mathcal{E}_t(h_t)).$$

In this paper, we consider two generalizations of the above result.

1) We allow $\mathcal{E}_t$ to be *any* estimator of the state rather than restricting attention to MMSE estimator.
2) We consider general POMDPs rather than restricting attention to linear systems.

In this general setting, we define the certainty equivalent policy $\mu^{\mathsf{CE}} = (\mu^{\mathsf{CE}}_1, \dots, \mu^{\mathsf{CE}}_T)$ as

$$\mu^{\mathsf{CE}}_t(h_t) = \pi^{\mathcal{M}}_t(\mathcal{E}_t(h_t)). \tag{1}$$

In general, μ^{CE} is not optimal. Our main result is to characterize the degree of sub-optimality of the certainty equivalent policy μ^{CE}.

Our results may be viewed as an instance of characterizing the sub-optimality gap of structured policies for POMDPs. There is a rich literature on deriving such sub-optimality gaps using tools from predictive state representation [28], [29], bisimulation metrics [30], approximation information states (AIS) [11], and filter stability [12]–[14]. Our analysis is based on AIS-based approximation bounds of [11].

The main contributions of this paper are as follows:

- We derive explicit bounds on the sub-optimality of certainty equivalent policies under the assumption that the cost and dynamics are smooth in an appropriate sense. Our bounds depend on the worst-case value of the conditional expected estimation error.
- We extend our results to settings with state abstraction, where the estimator produces an estimate of an abstract state rather than the full state, allowing our framework to apply to large-scale systems where state aggregation/quantization or feature abstraction is necessary.
- We illustrate our results through several examples, including: systems with bounded observation noise, intermittently degraded observations, control with event-triggered communication, learning and adaptive control settings, and control of non-homogeneous multi-particle systems. These examples demonstrate that certainty equivalent policies can achieve near-optimal performance when the estimation error is small, providing practical and computationally tractable alternatives to exact POMDP solutions.

A preliminary version of this result appeared in [1]. The analysis there was restricted to certainty equivalent policies which estimate the complete state of the system and the results were obtained under stronger assumptions. The state abstraction model considered in this paper is new and the results are derived under weaker assumptions.

The rest of the paper is organized as follows. We present the system model and define certainty equivalent policies in Sec. II. We illustrate our results through several examples in Sec. III. We present the proofs in Sec. IV and conclude in Sec. V.

Notation: We use uppercase letters to denote random variables (e.g., S, A, etc.), the corresponding lowercase letters to denote their realizations (e.g., s, a, etc.), and the corresponding calligraphic letters to denote their space of realizations (e.g., $\mathcal{S}$, $\mathcal{A}$, etc.). Subscripts denote time, so S_t denotes a variable at time t. The notation $S_{1:t}$ is a shorthand for the sequence $(S_1, \dots, S_t)$.

We use $\mathbb{R}$ to denote the set of real numbers. For a topological space $\mathcal{X}$, $\Delta(\mathcal{X})$ denotes the set of all probability measures on $\mathcal{X}$ and $\mathcal{B}(\mathcal{X})$ denotes the set of all bounded and measurable real-valued functions on $\mathcal{X}$.

We use $\mathbb{P}(\cdot)$ to denote the probability of an event and $\mathbb{E}[\cdot]$ to denote the expectation of a random variable. We use the notation $\mathbb{P}(S_{t+1} \in M_S | s_t, a_t)$ as a shorthand for $\mathbb{P}(S_{t+1} \in M_S | S_t = s_t, A_t = a_t)$.

Given a metric space $(\mathcal{S}, d_{\mathcal{S}})$, the Wasserstein-1 distance between two probability distributions $\nu_1, \nu_2 \in \Delta(\mathcal{S})$ is given by

$$d_{\mathrm{Was}}(\nu_1, \nu_2) = \inf_{(S_1, S_2) \sim \Gamma(\nu_1, \nu_2)} \mathbb{E}[d_{\mathcal{S}}(S_1, S_2)],$$

where $\Gamma(\nu_1, \nu_2)$ denotes all joint probability distributions on $\mathcal{S} \times \mathcal{S}$ with marginals ν_1 and ν_2. For two random variables S_1 and S_2 taking values in $\mathcal{S}$, we sometimes use $d_{\mathrm{Was}}(S_1, S_2)$ to denote the Wasserstein-1 distance between the marginal distributions of S_1 and S_2. For a function f from one metric space to another, $\mathrm{Lip}(f)$ denotes the Lipschitz constant of f.

II. System model and the main results

Consider a discrete-time partially observable Markov decision process (POMDP), denoted by $\mathcal{P}$, with state space $\mathcal{S}$, observation space $\mathcal{Y}$, and action space $\mathcal{A}$ that runs for a finite horizon T. Let $S_t \in \mathcal{S}$ denote the state of the system, $Y_t \in \mathcal{Y}$ denote the observation of the controller, and $A_t \in \mathcal{A}$ denote the control action taken by the controller at time t. We assume that $\mathcal{S}, \mathcal{Y}$ and $\mathcal{A}$ are metric spaces and use $d_{\mathcal{S}}$ to denote the metric on $\mathcal{S}$.

The initial state and observation (S_1, Y_1) are distributed according to a probability distribution $\xi \in \Delta(\mathcal{S} \times \mathcal{Y})$. The dynamics and observation are assumed to be Markovian. In particular, we assume that there exist stochastic kernels $P_t \colon \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S} \times \mathcal{Y})$, $t \in \{1, \dots, T-1\}$, such that for any $t \in \{1, \dots, T-1\}$, any Borel subsets M_S, M_Y of $\mathcal{S}$ and $\mathcal{Y}$ respectively, and any realizations $s_{1:t}$, $y_{1:t}$ and $a_{1:t}$ of $S_{1:t}$, $Y_{1:t}$, $A_{1:t}$, respectively, we have

$$\begin{aligned} &\mathbb{P}(S_{t+1} \in M_S, Y_{t+1} \in M_Y | s_{1:t}, y_{1:t}, a_{1:t}) \\ &\quad = \mathbb{P}(S_{t+1} \in M_S, Y_{t+1} \in M_Y | s_t, a_t) \\ &\quad =: P_t(M_S, M_Y | s_t, a_t). \end{aligned} \tag{2}$$

We will use the notation $P_{S,t}(\cdot | s_t, a_t)$ and $P_{Y,t}(\cdot | s_t, a_t)$ to denote the state and observation marginals of $P_t(\cdot, \cdot | s_t, a_t)$.

At each time t, the system incurs a per-step cost $c_t(S_t, A_t)$, which is uniformly bounded i.e., there exists a $c_{\max} \in \mathbb{R}$ such that $\sup_{s \in \mathcal{S}, a \in \mathcal{A}} |c_t(s, a)| \le c_{\max}$.

The controller has access to observation and action history $h_t = \{y_{1:t}, a_{1:t-1}\}$ at time t. Let $\mathcal{H}_t$ denote the space of realizations of h_t. Let $\mu = (\mu_1, \dots, \mu_T)$ denote any history

dependent deterministic policy. The value function of policy μ is defined as

$$W_t^{\mathcal{P},\mu}(h_t) = \mathbb{E}^{\mu}\left[\sum_{\tau=t}^{T} c_\tau(s_\tau, a_\tau) \;\middle|\; h_t\right]$$

where $\mathbb{E}^{\mu}$ denotes expectation with respect to a joint probability measure on the system variables induced by the policy μ. The *optimal* value function is defined as

$$W_t^{\mathcal{P}}(h_t) = \inf_{\mu} W_t^{\mathcal{P},\mu}(h_t),$$

where the infimum is over all history dependent policies.

The standard approach to find an optimal policy in POMDPs is to use belief-state based dynamic programs [2], [3], which are computationally challenging. As discussed in the introduction, certainty equivalent policies provide an attractive alternative approach. In the rest of this section, we characterize the sub-optimality of such policies.

A. Certainty equivalent policies

Consider a state feedback controller for the stochastic system defined above, where the controller has access to the state S_t at time t. This system is a finite horizon Markov decision process (MDP) $\mathcal{M}$ with state space $\mathcal{S}$, action space $\mathcal{A}$, dynamics $P_{S,t}$, and per-step cost c_t.

We need a technical assumption to ensure that the MDP $\mathcal{M}$ has an optimal policy.

Definition 1 (Measurable selection) An MDP $\langle \mathcal{S}, \mathcal{A}, \{P_{S,t}\}_{t=1}^{T-1}, \{c_t\}_{t=1}^{T}, T\rangle$ is said to satisfy *measurable selection* if for every measurable function $V\colon \mathcal{S} \to \mathbb{R}$ and each time $t \in \{1, \dots, T-1\}$, there exists a measurable selector $\pi\colon \mathcal{S} \to \mathcal{A}$ such that

$$\begin{aligned}&\inf_{a\in\mathcal{A}}\left\{c_t(s,a) + \int_{\mathcal{S}} V(s')P_{S,t}(ds'|s,a)\right\}\\ &= c_t(s,\pi(s)) + \int_{\mathcal{S}} V(s')P_{S,t}(ds'|s,\pi(s)) =: V_+(s),\end{aligned}$$

and $V_+\colon \mathcal{S} \to \mathbb{R}$ defined above is a measurable function.

Assumption 1 The model $\mathcal{M}$ satisfies measurable selection.

An implication of $\mathcal{M}$ satisfying measurable selection is that there exists an optimal policy $\pi^{\mathcal{M}} = (\pi_1^{\mathcal{M}}, \dots, \pi_T^{\mathcal{M}})$, where $\pi_t^{\mathcal{M}}\colon \mathcal{S} \to \mathcal{A}$, with associated optimal value functions $(V_1^{\mathcal{M}}, \dots, V_T^{\mathcal{M}})$, $V_t^{\mathcal{M}}\colon \mathcal{S} \to \mathbb{R}$, for this MDP [31].

We now use the optimal policy $\pi^{\mathcal{M}}$ for the MDP $\mathcal{M}$ to define a feasible policy for the POMDP $\mathcal{P}$. Suppose we are given a sequence of *state estimation functions* $\{\mathcal{E}_t\}_{t=1}^{T}$, where $\mathcal{E}_t\colon \mathcal{H}_t \to \mathcal{S}$. For instance, $\mathcal{E}_t$ may be the conditional mean or the MAP (maximum a posteriori probability) estimator which depend on the conditional distribution of the state given the history of observations and actions. Alternatively, the estimator could be a simple function (e.g. linear) of the last few observations and actions.

We say that a history-dependent policy $\mu^{\mathcal{E}} = (\mu_1^{\mathcal{E}}, \dots, \mu_T^{\mathcal{E}})$ is *certainty equivalent with respect to MDP* $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \{P_{S,t}\}_{t=1}^{T-1}, \{c_t\}_{t=1}^{T}, T\rangle$ *and estimators* $\{\mathcal{E}_t\colon \mathcal{H}_t \to \mathcal{S}\}_{t\geq 1}$ if

$$\mu_t^{\mathcal{E}}(h_t) = \pi_t^{\mathcal{M}}(\mathcal{E}_t(h_t)). \tag{3}$$

In other words, a certainty equivalent policy treats $\mathcal{E}_t(h_t)$ as an error-free estimate of the state of the MDP $\mathcal{M}$ and then acts according to the optimal policy of $\mathcal{M}$. As mentioned earlier, such policies are optimal in the LQG setting when the conditional mean is used as the estimate but they are, in general, not optimal. We are interested in providing a bound on the sub-optimality of the certainty equivalent policies. Specifically, we are interested in an upper bound on the gap between the value functions of policy $\mu^{\mathcal{E}}$ and the optimal value functions of the POMDP, i.e. a bound on $W_t^{\mathcal{P},\mu^{\mathcal{E}}}(h_t) - W_t^{\mathcal{P}}(h_t)$. Our results provide such a bound under the following technical assumption on the "smoothness" of per-step cost and system dynamics.

Assumption 2 There exist a sequence of concave and non-decreasing functions $F_t^P, F_t^c\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, $t \in \{1, \dots, T\}$, such that for any $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$, we have

$$d_{\mathrm{Was}}(P_{S,t}(\cdot|s,a), P_{S,t}(\cdot|s',a)) \leq F_t^P(d_{\mathcal{S}}(s,s')) \tag{4}$$

and

$$\big|c_t(s,a) - c_t(s',a)\big| \leq F_t^c(d_{\mathcal{S}}(s,s')). \tag{5}$$

Remark 1 When F_t^P, F_t^c are linear, i.e., $F_t^P(x) = L_t^P x$ and $F_t^c(x) = L_t^c x$ for some positive constants L_t^P and L_t^c, then Assumption 2 reduces to assuming that the dynamics and per-step cost are Lipschitz continuous, which is a standard assumption for smoothness of the dynamics and per-step cost, and implies smoothness (Lipschitz continuity) of the value function of MDP $\mathcal{M}$ [32]. In particular, following the argument of [32], we have

$$\mathrm{Lip}(V_t^{\mathcal{M}}) \leq L_t^c + L_{t+1}^P \,\mathrm{Lip}(V_{t+1}^{\mathcal{M}}),$$

which can be unrolled to obtain an upper bound on the Lipschitz constant of $V_t^{\mathcal{M}}$ in terms of $\{L_\tau^P\}_{\tau=t+1}^{T-1}$ and $\{L_\tau^c\}_{\tau=t}^{T}$.

Our bounds for the sub-optimality gap of certainty equivalent policy $\mu_t^{\mathcal{E}}(h_t)$ defined in (3) depend on the quality of the estimates produced by the state estimation functions $\mathcal{E}_t$, which we assess using the metric $d_{\mathcal{S}}$ on the state space. For each time t, we define

$$\eta_t := \sup_{h_t \in \mathcal{H}_t} \mathbb{E}[d_{\mathcal{S}}(S_t, \mathcal{E}_t(h_t))|h_t]. \tag{6}$$

Remark 2 Note that in Eq. (6) the right hand side does not depend on the policy because the conditional probability distribution of current state S_t given history h_t is policy independent [2], [3].

Assumption 3 For $t = 1, 2, \dots, T$ we have $\eta_t < \infty$ where η_t is given by (6).

We can now state our first result.

Theorem 1 *Define*

$$\varepsilon_t = F_t^c(\eta_t) \text{ and } \delta_t = F_t^P(\eta_t) + \eta_{t+1}$$

where η_t is given by (6). Then, under Assumptions 1, 2 and 3, we have that the certainty equivalent policy $\mu^{\mathcal{E}}$ (defined in (3)) satisfies

$$W_t^{\mathcal{P},\mu^{\mathcal{E}}}(h_t) - W_t^{\mathcal{P}}(h_t) \le 2\alpha_t \tag{7}$$

where

$$\alpha_t = \varepsilon_t + \sum_{\tau=t}^{T-1} \left[\delta_\tau \operatorname{Lip}(V_{\tau+1}^{\mathcal{M}}) + \varepsilon_{\tau+1}\right] \tag{8}$$

and $\{V_t^{\mathcal{M}}\}_{t=1}^T$ are the optimal value functions for MDP $\mathcal{M}$.

We will state and prove a more general version of this result in the next subsection.

Remark 3 Certainty equivalent policies are not appropriate for all models. For instance, in some POMDPs, before taking a control action, the agent has the option to pay a cost to take a sensing action that reveals the true state of the MDP. In such models, the certainty equivalent policy will never choose the sensing action. Therefore, if the sensing action is not too costly, a policy that occasionally pays the sensing cost to learn the true state may outperform certainty equivalent policy.

B. Certainty equivalent policies using State Abstraction

In problems with large or continuous state spaces, it can be difficult to find an optimal policy $\pi^{\mathcal{M}}$ of MDP $\mathcal{M}$ due to the curse of dimensionality. For such large-scale MDPs, one typically obtains an approximately optimal policy for MDP $\mathcal{M}$ by considering an abstract MDP obtained by state aggregation or state quantization. In such situations, it is natural to consider a certainty equivalent policy based on an optimal policy of the abstract MDP and estimates of the abstract state. In this section, we formally define such a policy and present a generalization of Theorem 1 to such settings.

Suppose there is an abstract state space $\tilde{\mathcal{S}}$ which is equipped with a metric $d_{\tilde{\mathcal{S}}}$, a (measurable) state abstraction function $\phi\colon \mathcal{S} \to \tilde{\mathcal{S}}$, and two stochastic kernels $\lambda^P, \lambda^c\colon \tilde{\mathcal{S}} \to \Delta(\mathcal{S})$ such that for each $\tilde{s}_t \in \tilde{\mathcal{S}}$, $\lambda^P(\phi^{-1}(\tilde{s}_t)|\tilde{s}_t) = 1$ and $\lambda^c(\phi^{-1}(\tilde{s}_t)|\tilde{s}_t) = 1$.

We construct an abstract MDP $\tilde{\mathcal{M}} := \langle \tilde{\mathcal{S}}, \mathcal{A}, \{\tilde{P}_t\}_{t=1}^{T-1}, \{\tilde{c}_t\}_{t=1}^T, T\rangle$ where the dynamics $\tilde{P}_t\colon \tilde{\mathcal{S}} \times \mathcal{A} \to \tilde{\mathcal{S}}$ and the per-step cost $\tilde{c}_t\colon \tilde{\mathcal{S}} \times \mathcal{A} \to \mathbb{R}$ are defined as follows: for any measurable $M_{\tilde{\mathcal{S}}} \subset \tilde{\mathcal{S}}$,

$$\begin{aligned}\tilde{P}_t(\tilde{S}_{t+1} \in & M_{\tilde{\mathcal{S}}}|\tilde{s}_t, a_t) \\ &= \int_{\phi^{-1}(\tilde{s}_t)} P_{\mathcal{S},t}\big(\phi(S_{t+1}) \in M_{\tilde{\mathcal{S}}}|s_t, a_t)\big)\lambda^P(ds_t|\tilde{s}_t)\end{aligned} \tag{9}$$

and

$$\tilde{c}_t(\tilde{s}_t, a_t) = \int_{\phi^{-1}(\tilde{s}_t)} c_t(s_t, a_t)\lambda^c(ds_t|\tilde{s}_t). \tag{10}$$

The cost function of the abstract MDP can be viewed as a weighted averaging of the original MDP cost over all states in $\phi^{-1}(\tilde{s}_t)$; a similar interpretation applies for the dynamics in the abstract model as well.

Note that when $\tilde{\mathcal{S}} = \mathcal{S}$ and $\phi(s) = s$, the abstract MDP $\tilde{\mathcal{M}}$ is equal to the MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \{P_{S,t}\}_{t=1}^{T-1}, \{c_t\}_{t=1}^T, T\rangle$.

We impose the measurable selection assumption on MDP $\tilde{\mathcal{M}}$.

Assumption 4 The model $\tilde{\mathcal{M}}$ satisfies measurable selection.

As in Section II-A, an implication of measurable selection is that there exists an optimal policy $\pi^{\tilde{\mathcal{M}}} = (\pi_1^{\tilde{\mathcal{M}}}, \dots, \pi_T^{\tilde{\mathcal{M}}})$, where $\pi_t^{\tilde{\mathcal{M}}}\colon \tilde{\mathcal{S}} \to \mathcal{A}$, with associated optimal value functions $(V_1^{\tilde{\mathcal{M}}}, \dots, V_T^{\tilde{\mathcal{M}}})$, $V_t^{\tilde{\mathcal{M}}}\colon \tilde{\mathcal{S}} \to \mathbb{R}$, for MDP $\tilde{\mathcal{M}}$ [31].

We now use the optimal policy $\pi^{\tilde{\mathcal{M}}}$ for the MDP $\tilde{\mathcal{M}}$ to define a feasible policy for the POMDP $\mathcal{P}$. This policy is similar to the certainty equivalent policies of Section II-A except that (i) we estimate the abstract state $\tilde{S}_t$, and (ii) use an optimal policy of $\tilde{\mathcal{M}}$. For convenience, we reuse some of the notation of Section II-A.

Suppose we are given a sequence of *abstract state estimation functions* $\{\mathcal{E}_t\}_{t=1}^T$, where $\mathcal{E}_t\colon \mathcal{H}_t \to \tilde{\mathcal{S}}$. We say that a history-dependent policy $\mu^{\mathcal{E}} = (\mu_1^{\mathcal{E}}, \dots, \mu_T^{\mathcal{E}})$ is *certainty equivalent* with respect to the abstract MDP $\tilde{\mathcal{M}} = \langle \tilde{\mathcal{S}}, \mathcal{A}, \{\tilde{P}_t\}_{t=1}^{T-1}, \{\tilde{c}_t\}_{t=1}^T, T\rangle$ and estimators $\{\mathcal{E}_t\colon \mathcal{H}_t \to \tilde{\mathcal{S}}\}_{t\ge1}$ if

$$\mu_t^{\mathcal{E}}(h_t) = \pi_t^{\tilde{\mathcal{M}}}(\mathcal{E}_t(h_t)). \tag{11}$$

In other words, a certainty equivalent policy treats $\mathcal{E}_t(h_t)$ as an error-free estimate of the state of the abstract MDP and uses the optimal policy of $\tilde{\mathcal{M}}$ to take its action. As before, we are interested in an upper bound on the gap between the value functions of policy $\mu^{\mathcal{E}}$ and the optimal value functions of the POMDP, i.e. a bound on $W_t^{\mathcal{P},\mu^{\mathcal{E}}}(h_t) - W_t^{\mathcal{P}}(h_t)$.

We impose the following technical assumption on the model and the state abstraction.

Assumption 5 There exist a sequence of concave and non-decreasing functions $F_t^P, F_t^c\colon \mathbb{R}_{\ge0} \to \mathbb{R}_{\ge0}$, $t \in \{1, \dots, T\}$, such that for any $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$, we have

$$d_{\mathrm{Was}}(P_{S,t}^{\phi}(\cdot|s,a), P_{S,t}^{\phi}(\cdot|s',a)) \le F_t^P(d_{\tilde{\mathcal{S}}}(\phi(s),\phi(s'))), \tag{12}$$

and

$$\big|c_t(s,a) - c_t(s',a)\big| \le F_t^c(d_{\tilde{\mathcal{S}}}(\phi(s),\phi(s'))) \tag{13}$$

where $P_{S,t}^{\phi}$ is a stochastic kernel from $\mathcal{S}\times\mathcal{A}$ to $\Delta(\tilde{\mathcal{S}})$ defined as

$$P_{S,t}^{\phi}(M_{\tilde{\mathcal{S}}}|s_t,a_t) = P_{\mathcal{S},t}\big(\phi(S_{t+1}) \in M_{\tilde{\mathcal{S}}}|s_t,a_t\big)$$

for all Borel subsets $M_{\tilde{\mathcal{S}}}$ of $\tilde{\mathcal{S}}$.

Assumption 5 implies that $\phi : \mathcal{S} \to \tilde{\mathcal{S}}$ is a good state abstraction in the following sense: If $\phi(s)$ is close to $\phi(s')$, then for any action a, the per-step costs at s and s' are close and the probability distributions of the next abstracted state given s and s' are close.

Our results depend on the quality of the estimates produced by $\mathcal{E}_t$. For that purpose, for each time t we define the worst-case value of the conditional expected estimation error

$$\eta_t := \sup_{h_t\in\mathcal{H}_t} \mathbb{E}[d_{\tilde{\mathcal{S}}}(\phi(S_t), \mathcal{E}_t(h_t))|h_t]. \tag{14}$$

As we stated in Remark 2, the right hand side in Eq. (14) does not depend on the policy because conditional probability

distribution of the state S_t given history h_t is policy independent [2], [3].

Assumption 6 For $t = 1, 2, \dots, T$, we have $\eta_t < \infty$ where η_t is given by (14).

We can now state the main result.

Theorem 2 *Define*

$$\varepsilon_t = F_t^c(\eta_t) \text{ and } \delta_t = F_t^P(\eta_t) + \eta_{t+1}$$

where η_t is given by (14). Then, under Assumptions 4, 5 and 6, we have that the certainty equivalent policy $\mu^{\mathcal{E}}$ (defined in (11)) satisfies

$$W_t^{\mathcal{P},\mu^{\mathcal{E}}}(h_t) - W_t^{\mathcal{P}}(h_t) \le 2\alpha_t \tag{15}$$

where

$$\alpha_t = \varepsilon_t + \sum_{\tau=t}^{T-1} \left[\delta_\tau \operatorname{Lip}(V_{\tau+1}^{\tilde{\mathcal{M}}}) + \varepsilon_{\tau+1}\right] \tag{16}$$

and $\{V_t^{\tilde{\mathcal{M}}}\}_{t=1}^T$ are the optimal value functions for MDP $\tilde{\mathcal{M}}$.

We will prove this result in Section IV. Note that when $\tilde{\mathcal{S}} = \mathcal{S}$ and $\phi(s) = s$, Theorem 2 reduces to Theorem 1. As illustrated in the corollary below, Theorem 2 also provide sub-optimality bounds for using a policy of an abstract MDP in the original MDP, which may be viewed as a finite horizon version of [33].

Corollary 1 *For any Markovian policy π of MDP $\mathcal{M}$, let $V_t^{\mathcal{M},\pi} \colon \mathcal{S} \to \mathbb{R}$ denote the value function of policy π. Given the optimal policy $\pi^{\tilde{\mathcal{M}}} = (\pi_1^{\tilde{\mathcal{M}}}, \dots, \pi_T^{\tilde{\mathcal{M}}})$ of the abstract MDP $\tilde{\mathcal{M}}$, define a feasible policy $\bar{\pi}$ of MDP $\mathcal{M}$ as follows: $\bar{\pi} = (\pi_1^{\tilde{\mathcal{M}}} \circ \phi, \dots, \pi_T^{\tilde{\mathcal{M}}} \circ \phi)$. Then, under Assumptions 4 and 5, for any time t and any realization s_t of S_t, we have*

$$V_t^{\mathcal{M},\bar{\pi}}(s_t) - V^{\mathcal{M}}(s_t) \le 2\alpha_t$$

where α_t is given by (16) with $\varepsilon_t = F_t^c(0)$ and $\delta_t = F_t^P(0)$.

PROOF This is an immediate consequence of Theorem 2 by considering the trivial setting where $Y_t = S_t$ (and thus, $h_t = (s_{1:t}, a_{1:t-1})$) and take $\mathcal{E}(h_t) = \phi(s_t)$, which implies $\eta_t = 0$ for all t. In this case, $W_t^{\mathcal{P}}(h_t) = V_t^{\mathcal{M}}(s_t), \mu^{\mathcal{E}} = \bar{\pi}$ and $W_t^{\mathcal{P},\mu^{\mathcal{E}}}(h_t) = V_t^{\mathcal{M},\bar{\pi}}(s_t)$. ∎

III. Illustrative examples

In this section we present several examples to illustrate observation models (and corresponding estimators) where certainty equivalent policies may be useful. We apply our results to derive explicit bounds on the sub-optimality of certainty equivalent policies for specific observation models.

A. Bounded observation noise

1) System model: Consider a POMDP where $\mathcal{Y} = \mathcal{S}$ and the system dynamics P_t are such that

$$d_{\mathcal{S}}(Y_t, S_t) \le r,$$

where $r \in [0, \infty)$. Moreover, we assume that the MDP model $\mathcal{M}$ satisfies measurable selection (Assumption 1) and that the dynamics and per-step cost are Lipschitz, i.e., there exist non-negative finite constants L_t^P and L_t^c such that Assumption 2 is satisfied with $F_t^P(x) = L_t^P x$ and $F_t^c(x) = L_t^c x$ (see Remark 1).

2) Certainty equivalent policy: For this example, we consider certainty equivalent policies with respect to the original MDP $\mathcal{M}$, i.e., take the state abstraction function $\phi(s) = s$. Furthermore, we take the state estimate to be the last observation, i.e., $\mathcal{E}_t(h_t) = y_t$. Then, the certainty equivalent policy is given by

$$\mu_t^{\mathcal{E}}(h_t) = \pi_t^{\mathcal{M}}(y_t).$$

3) Sub-optimality bound: We have assumed that Assumptions 1 and 2 are satisfied. Moreover,

$$\mathbb{E}[d_{\mathcal{S}}(S_t, \mathcal{E}_t(H_t))|h_t] = \mathbb{E}\big[d_{\mathcal{S}}(S_t, Y_t)|h_t\big] \le r.$$

Therefore, $\eta_t \le r$ and Assumption 3 is also satisfied. Furthermore, the ε_t and δ_t in Theorem 1 can be upper bounded by

$$\varepsilon_t \le r L_t^c \quad \text{and} \quad \delta_t \le r(1 + L_t^P).$$

Hence, the bound in Theorem 1 can be explicitly written as

$$W_t^{\mathcal{P},\mu^{\mathcal{E}}}(h_t) - W_t^{\mathcal{P}}(h_t) \le 2r L_t^{\mathcal{M}} \tag{17}$$

where

$$L_t^{\mathcal{M}} = \left[L_t^c + \sum_{\tau=t}^{T-1} \left[(1 + L_\tau^P) \operatorname{Lip}(V_{\tau+1}^{\mathcal{M}}) + L_{\tau+1}^c\right]\right] \tag{18}$$

is a constant that depends on the Lipschitz constants of the dynamics, per-step cost, and the optimal MDP value function.

This bound scales linearly with r, which means that as the observation becomes closer to the underlying state (i.e., "observation noise" becomes small), the performance of the certainty equivalent policy approaches that of the optimal POMDP policy.

B. Intermittently degraded observation

1) System model: Consider a POMDP where $\mathcal{Y} = \mathcal{S}$ and the system dynamics P_t is such that the controller either gets a good observation (indicated by event E_t) or a bad observation (indicated by E_t^c). Under E_t, $d_{\mathcal{S}}(Y_t, S_t) \le r$ while under E_t^c, $d_{\mathcal{S}}(Y_t, S_t) \le R$ where $0 \le r \le R < \infty$. We assume that for any history h_t, $\mathbb{P}(E_t^c|h_t) \le p$.

Moreover, we assume that the MDP model $\mathcal{M}$ satisfies measurable selection (Assumption 1) and that the dynamics and per-step cost are Lipschitz, i.e., there exist non-negative finite constants L_t^P and L_t^c such that Assumption 2 is satisfied with $F_t^P(x) = L_t^P x$ and $F_t^c(x) = L_t^c x$ (see Remark 1).

2) Certainty equivalent policy: As for the example in Sec. III-A, we take $\phi(s) = s$ and $\mathcal{E}_t(h_t) = y_t$. Then, the certainty equivalent policy is given by

$$\mu_t^{\mathcal{E}}(h_t) = \pi_t^{\mathcal{M}}(y_t).$$

3) Sub-optimality bound: We have assumed that Assumption 1 and 2 are satisfied. Moreover,

$$\begin{aligned}\mathbb{E}[d_{\mathcal{S}}(S_t, \mathcal{E}_t(H_t))|h_t] &= \mathbb{E}\big[d_{\mathcal{S}}(S_t, Y_t)|h_t\big] \\ &\le (1-p)r + pR \end{aligned} \tag{19}$$

Hence, we have $\eta_t \le (1-p)r + pR$ and Assumption 3 is also satisfied. Furthermore, the ε_t and δ_t in Theorem 1 can be upper bounded by

$$\varepsilon_t \le [(1-p)r + pR] L_t^c \quad \text{and} \quad \delta_t \le [(1-p)r + pR](1 + L_t^P).$$

Therefore, the bound in Theorem 1 can be explicitly written as

$$W_t^{\mathcal{P},\mu^{\mathcal{E}}}(h_t) - W_t^{\mathcal{P}}(h_t) \le 2[(1-p)r + pR] L_t^{\mathcal{M}} \tag{20}$$

where $L_t^{\mathcal{M}}$ is as defined in (18).

These results demonstrate that when the state estimation error $[(1-p)r + pR]$ is small, either due to observations with small noise or observations that are frequently accurate, certainty equivalent policies can perform near-optimally. The bounds provide a quantitative measure of the sub-optimality in terms of the estimation error and the Lipschitz constants of the model.

C. Bounded observation noise with state quantization

1) System model: Consider the bounded observation noise model of Sec. III-A where $\mathcal{Y} = \mathcal{S}$ and the observation model is such that $d_{\mathcal{S}}(Y_t, S_t) \le r$. In addition, an abstract model $\hat{\mathcal{M}}$ is given, which is constructed by partitioning the state space $\mathcal{S}$ into a finite collection $\{\Psi_k\}_{k=1}^K$ of quantization cells, each with a representative element $\tilde{s}_k \in \Psi_k$. The abstract state space is $\tilde{\mathcal{S}} = \{\tilde{s}_1, \dots, \tilde{s}_K\}$ and the state abstraction function is $\phi : \mathcal{S} \to \tilde{\mathcal{S}}$ given by $\phi(s) = \tilde{s}_k$ for all $s \in \Psi_k$, $k \in \{1, \dots, K\}$. Moreover, the stochastic kernels $\lambda^P(\cdot \mid \tilde{s}_k)$ and $\lambda^c(\cdot \mid \tilde{s}_k)$ are Dirac delta measures on $\tilde{s}_k$, $k \in \{1, \dots, K\}$. The abstract dynamics $\tilde{P}$ and abstract cost $\tilde{c}$ are constructed as in (9) and (10). We also define a metric $d_{\tilde{\mathcal{S}}}$ on $\tilde{\mathcal{S}}$ by

$$d_{\tilde{\mathcal{S}}}(\tilde{s}_1, \tilde{s}_2) \coloneqq d_{\mathcal{S}}(\tilde{s}_1, \tilde{s}_2), \quad \forall \tilde{s}_1, \tilde{s}_2 \in \tilde{\mathcal{S}} \subset \mathcal{S},$$

We assume Assumptions 4 and 5 are satisfied.

2) Certainty equivalent policy: For this example, we consider certainty equivalent policies with respect to the abstract MDP $\tilde{\mathcal{M}}$. Furthermore, we take the abstract state estimate to be the quantized value of the last observation, i.e., $\mathcal{E}_t(h_t) = \phi(y_t)$. Then, the certainty equivalent policy is

$$\mu_t^{\mathcal{E}}(h_t) = \pi^{\tilde{M}}\big(\phi(y_t)\big).$$

3) Sub-optimality bound: We have assumed that Assumption 4 and 5 are satisfied. Let R denote the maximum radius of a quantization cell, i.e.,

$$R \coloneqq \max_{k \in \{1,\dots,K\}} \sup_{s \in \Psi_k} d_{\mathcal{S}}(s, \phi(s)).$$

Then, by triangle inequality, we have

$$\begin{aligned}
&\mathbb{E}[d_{\tilde{\mathcal{S}}}(\phi(S_t), \mathcal{E}_t(h_t)) | h_t] = \mathbb{E}[d_{\mathcal{S}}(\phi(S_t), \phi(Y_t) | h_t] \\
&\le \mathbb{E}\big[d_{\mathcal{S}}\big(\phi(S_t), S_t\big) + d_{\mathcal{S}}(S_t, Y_t) + d_{\mathcal{S}}\big(Y_t, \phi(Y_t)\big) | h_t\big] \\
&\le R + r + R = r + 2R.
\end{aligned} \tag{21}$$

Hence, we have $\eta_t \le r + 2R$ and Assumption 6 is also satisfied. Furthermore, the ε_t and δ_t in Theorem 2 can be upper bounded by

$$\varepsilon_t \le F_t^c(r + 2R) \quad \text{and} \quad \delta_t \le F_t^P(r + 2R) + r + 2R.$$

Therefore, we can upper bound the sub-optimality of using the certainty equivalent policy by the bound (15) in Theorem 2.

D. Certainty equivalence in learning/adaptive control

1) System model: Consider a parameterized MDP $\mathcal{M}_X(\theta)$ with state space $\mathcal{X}$, action space $\mathcal{A}$, time-invariant dynamics $P_{X,\theta}$ and time-invariant per-step cost ℓ_θ, where the parameters $\theta \in \Theta$ and are distributed according to some probability distribution P_Θ independent of noise in the dynamics. We assume that $\mathcal{X}$ and Θ are metric spaces with metrics $d_{\mathcal{X}}$ and d_Θ, respectively.

The controller doesn't know the parameters θ but knows the history of state and actions $h_t = (x_{1:t}, a_{1:t-1})$. The above model can be viewed as an POMDP with state space $\mathcal{S} = \mathcal{X} \times \Theta$, observation space $\mathcal{X} \times \mathbb{R}$, where $S_t = (X_t, \theta)$ and $Y_t = (X_t, \ell(X_{t-1}, A_{t-1}))$. We take $d_{\mathcal{S}}((x_1, \theta_1), (x_2, \theta_2)) = d_{\mathcal{X}}(x_1, x_2) + d_\Theta(\theta_1, \theta_2)$. Observe that the MDP $\mathcal{M}$ corresponding to this POMDP is equivalent to $\mathcal{M}_X(\theta)$. We denote the optimal policy of this MDP by $\pi^{\mathcal{M}}(x, \theta) = \pi^{\mathcal{M}_X(\theta)}(x)$.

We assume that for all $\theta \in \Theta$, the model $\mathcal{M}_X(\theta)$ satisfies measurable selection (Assumption 1). Moreover, there exist non-negative finite constants L^P and L^c such that for any $x, x' \in \mathcal{X}$ and $\theta, \theta' \in \Theta$, we have

$$\begin{aligned}
&d_{\mathrm{Was}}(P_{X,\theta}(\cdot|x,a), P_{X,\theta'}(\cdot|x',a)) \\
&\qquad \le L^P(d_{\mathcal{X}}(x, x') + d_\Theta(\theta, \theta')),
\end{aligned} \tag{22}$$

and

$$\big|\ell_\theta(x,a) - \ell_{\theta'}(x',a)\big| \le L^c(d_{\mathcal{X}}(x, x') + d_\Theta(\theta, \theta')). \tag{23}$$

2) Certainty equivalent policy: As in the previous examples, we consider $\phi(s) = s$ but consider a general estimator $\mathcal{E}_t(h_t) = (x_t, \hat{\theta}_t)$ where $\hat{\theta}_t$ is some estimate of θ based on h_t, e.g., the MMSE estimator $\hat{\theta}_t = \mathbb{E}[\theta | h_t]$. Then, the certainty equivalent policy is

$$\mu_t^{\mathcal{E}}(h_t) = \pi_t^{\mathcal{M}}(x_t, \hat{\theta}_t).$$

3) Sub-optimality bound: We have assumed that Assumption 1 and 2 are satisfied. Moreover,

$$\eta_t = \sup_{h_t \in \mathcal{H}_t} \mathbb{E}[d_{\mathcal{S}}(S_t, \mathcal{E}_t(H_t)) | h_t] = \sup_{h_t \in \mathcal{H}_t} \mathbb{E}[d_\Theta(\theta_t, \hat{\theta}_t) | h_t]. \tag{24}$$

Thus, if Assumption 3 holds, the ε_t and δ_t in Theorem 1 can be upper bounded by

$$\varepsilon_t \le L^c \eta_t \quad \text{and} \quad \delta_t \le L^P \eta_t + \eta_{t+1}.$$

Therefore, we can bound the sub-optimality in using the certainty equivalent policy by the bound (7) given in Theorem 1.

These results show that the performance of certainty equivalent policies depends on the performance η_t of the parameter estimation. If η_t decays sufficiently fast, e.g., exponentially fast, then we can obtain uniform upper bounds on the performance error $2\alpha_t$ in (7) even when $T \to \infty$.

E. Control with event-triggered communication

1) System model: Consider a system consisting of a plant, a sensor co-located with the plant, and a remote controller. Let $X_t \in \mathcal{X}$ and $A_t \in \mathcal{A}$ denote the state and control input of the plant. The state of the plant evolves according to a controlled transition kernel $P_{X,t} \colon \mathcal{X} \times \mathcal{A} \to \Delta(\mathcal{X})$.

The sensor observes the current state X_t and decides whether or not to transmit the state. Let $Y_t \in \mathcal{X} \cup \{\mathfrak{E}\}$ denote the observation of remote controller, i.e.,

$$Y_t = \begin{cases} \mathfrak{E} & \text{if sensor does not communicate} \\ X_t & \text{if sensor communicates} \end{cases}$$

where $\mathfrak{E}$ denotes a null observation.

The remote controller generates the action A_t according to a general history dependent policy $\mu = (\mu_1, \dots, \mu_T)$ and incurs a per-step cost $c_t(x_t, a_t)$.

The above problem is a decentralized control problem where both the communication policy at the sensor and the control policy at the remote controller have to be determined. We consider the setting when the communication policy is fixed to be an *event-triggered* communication policy [34]–[36], which operates as follows. It is assumed that the remote controller keeps track of a state estimate $\hat{X}_{t|t} \in \mathcal{X}$ as follows:

$$\hat{X}_{t|t} = \begin{cases} Y_t & \text{if } Y_t \neq \mathfrak{E} \\ \hat{X}_{t|t-1}, & \text{if } Y_t = \mathfrak{E} \end{cases} \tag{25}$$

where $\hat{X}_{1|0} = \mathbb{E}[X_1]$ and

$$\hat{X}_{t|t-1} = g(\hat{X}_{t-1|t-1}, A_{t-1}), \quad t > 1, \tag{26}$$

where $g \colon \mathcal{X} \times \mathcal{A} \to \mathcal{X}$ is a pre-specified update function.

In an event-triggered policy, the sensor transmits when the following inequality is satisfied

$$d_{\mathcal{X}}(X_t, \hat{X}_{t|t-1}) > r.$$

where r is a pre-specified constant. This policy ensure that $d_{\mathcal{X}}(X_t, \hat{X}_{t|t}) \leq r$. The objective then is to find the best control strategy at the remote controller.

Once the event triggered policy is fixed, the above model corresponds to a POMDP $\mathcal{P}$ with state $S_t = (X_t, \hat{X}_{t|t-1})$, observation Y_t, and action A_t. Let $\mathcal{M}$ denote the corresponding MDP, in which the controller has access to S_t.

We consider an abstract MDP $\tilde{\mathcal{M}}$ with state space $\mathcal{X}$ which is constructed using a state abstraction function $\phi(x, \hat{x}) = x$ and stochastic kernels $\lambda^P(\cdot|x)$ and $\lambda^c(\cdot|x)$ as Dirac delta measures on (x, x). Then, the dynamics $\tilde{P}_t$ of the abstract MDP is equal to $P_{X,t}$ and the per-step cost $\tilde{c}_t$ is equal to c_t. Thus, MDP $\tilde{M} = \langle \mathcal{X}, \mathcal{A}, \{P_{X,t}\}_{t=1}^{T-1}, \{c_t\}_{t=1}^{T} \rangle$. We assume that $\tilde{\mathcal{M}}$ satisfies Assumption 4 and there exist concave and non-decreasing functions $F_t^P, F_t^c \colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, $t \in \{1, \dots, T\}$, such that for any $x, x' \in \mathcal{X}$ and $a \in \mathcal{A}$, we have

$$d_{\text{Was}}(P_{X,t}(\cdot|x, a), P_{X,t}(\cdot|x', a)) \leq F_t^P(d_{\mathcal{X}}(x, x')) \tag{27}$$

and

$$\big|c_t(x, a) - c_t(x', a)\big| \leq F_t^c(d_{\mathcal{X}}(x, x')). \tag{28}$$

Assumption 4 implies that there exists an optimal policy for MDP $\tilde{\mathcal{M}}$, which we denote by $\pi^{\tilde{\mathcal{M}}}$. Moreover, it can be verified that (27) and (28) implies Assumption 5.

2) Certainty equivalent policy: For this example, we consider certainty equivalent policies with respect to the abstract MDP $\tilde{\mathcal{M}}$. Furthermore, we take the state estimate to be $\mathcal{E}_t(h_t) = \hat{x}_{t|t}$, which is recursively computed from the history using (25). Then, the certainty equivalent policy is given by

$$\mu_t^{\mathcal{E}}(h_t) = \pi_t^{\tilde{\mathcal{M}}}(\hat{x}_{t|t}).$$

3) Sub-optimality bound: We have assumed Assumption 4 and assumed sufficient conditions that imply Assumption 5. Recall that the event-triggered sensor transmission policy ensures that $d_{\mathcal{X}}(X_t, \hat{X}_{t|t}) \leq r$. Therefore,

$$\mathbb{E}[d_{\mathcal{X}}(\phi(S_t), \mathcal{E}(H_t))|h_t] = \mathbb{E}[d_{\mathcal{X}}(X_t, \hat{X}_{t|t})|h_t] \leq r.$$

Hence, $\eta_t \leq r$ and Assumption 6 is satisfied. Furthermore, the ε_t and δ_t in Theorem 2 can be upper bounded by

$$\varepsilon_t \leq F_t^c(r) \quad \text{and} \quad \delta_t \leq F_t^P(r) + r.$$

Therefore, we can upper bound the sub-optimality of using the certainty equivalent policy by the bound (15) in Theorem 2. These bounds quantify the sub-optimality gap of certainty equivalent control with event-triggered sensing.

F. Control of non-homogeneous multi-particle systems

1) System model: Consider a system consisting of n particles, where each particle i, $i \in \{1, \dots, n\}$, has a state $X_t^i \in \mathbb{R}$. The global state of the system is $X_t = (X_t^1, X_t^2, \dots, X_t^n)^\intercal \in \mathbb{R}^n$. The observation is given by

$$Y_t = X_t + N_t$$

where $N_t = (N_t^1, N_t^2, \dots, N_t^n)^\intercal$ is the observation noise with $N_t^i \in [-r^i, r^i]$ for some $r^i \geq 0$.

Let $M_t = \sum_{i=1}^n \alpha^i X_t^i$ denote the weighted mean of the global state, where α^i, $i \in \{1, \dots, n\}$, are non-negative weights that add to 1. We assume that the dynamics of each particle is given by

$$X_{t+1}^i = \bar{f}(M_t, A_t, W_t) + f^i(X_t, A_t, W_t) \tag{29}$$

where $A_t \in \mathcal{A}$ is the control action at time t and $\{W_t\}_{t \geq 1}$, $W_t \in \mathcal{W}$, is an independent and identically distributed process with distribution P_W.

Assumption 7 We assume the following.

1) There exists a $L^{\bar{f}} \in [0, \infty)$ such that for all $m, m' \in \mathbb{R}$ and $a \in \mathcal{A}$,

$$d_{\text{Was}}(Z, Z') \leq L^{\bar{f}} |m - m'|. \tag{30}$$

where $Z \coloneqq f(m, a, W_t)$ and $Z' \coloneqq f(m', a, W_t)$.

2) There exists constants γ^i such that $\|f^i\|_\infty \leq \gamma^i$ for $i \in \{1, \dots, n\}$.

The per-step cost is given by

$$c(X_t, A_t) = \bar{\ell}(M_t, A_t) + \ell(X_t, A_t). \tag{31}$$

Assumption 8 We assume the following.

1) The function $\bar{\ell}$ is $L^{\bar{\ell}}$-Lipschitz, i.e. for all $m, m' \in \mathbb{R}$ and $a \in \mathcal{A}$,

$$|\bar{\ell}(m, a) - \bar{\ell}(m', a)| \leq L^{\bar{\ell}} |m - m'|. \tag{32}$$

2) There exists a constant β such that $\|\ell\|_\infty \leq \beta$.

The above model is a POMDP $\mathcal{P}$ with state $S_t = X_t$, observation Y_t, action A_t, and per-step cost $c(X_t, A_t)$.

2) Certainty equivalent policy: We consider an abstract MDP that focuses on the weighted mean of the global state. The abstract state space is $\tilde{\mathcal{S}} = \mathbb{R}$ and the state abstraction function $\phi : \mathbb{R}^n \to \mathbb{R}$ is given by

$$\phi((x^1, x^2, \dots, x^n)^\intercal) = \sum_{i=1}^n \alpha^i x^i$$

with

$$\phi^{-1}(m) = \Big\{(x^1, x^2, \dots, x^n)^\intercal : \sum_{i=1}^n \alpha^i x^i = m\Big\}.$$

We assume that the metric on the abstract state space is $d_{\tilde{\mathcal{S}}}(m, m') = |m - m'|$ and take $\lambda^P(\cdot|m)$ and $\lambda^c(\cdot|m)$ to be delta distributions at $m\mathbf{1}_n$, where $\mathbf{1}_n$ is the n-dimensional vector of ones.

Therefore, the abstract per-step cost is given by

$$\tilde{c}(\tilde{s}, a) = c(\tilde{s}\mathbf{1}_n, a) = \bar{\ell}(\tilde{s}, a) + \ell(\tilde{s}\mathbf{1}_n),$$

and the abstract state dynamics are given by

$$\tilde{S}_{t+1} = \bar{f}(\tilde{S}_t, A_t, W_t) + \sum_{i=1}^n \alpha^i f^i(\tilde{S}_t\mathbf{1}_n, A_t, W_t).$$

The above model corresponds to an abstract MDP $\tilde{\mathcal{M}}$. We assume that $\tilde{\mathcal{M}}$ satisfies Assumption 4.

We consider the weighted mean of the last observation as an estimate of the abstract state, i.e.,

$$\mathcal{E}_t(h_t) = \sum_{i=1}^n \alpha^i y_t^i.$$

Then, the certainty equivalent policy is

$$\mu_t^{\mathcal{E}}(h_t) = \pi_t^{\tilde{\mathcal{M}}}(\mathcal{E}(h_t)).$$

3) Sub-optimality bound: We have assumed that the abstract model $\tilde{\mathcal{M}}$ satisfies Assumption 4. We now show that Assumption 5 is satisfied.

Lemma 1 *Assumptions 7 and 8 imply Assumption 5 is satisfied with F_t^c and F_t^P defined as*

$$F_t^c(r) = L^{\bar{\ell}} r + 2\beta \quad \text{and} \quad F_t^P(r) = L^{\bar{f}} r + 2\sum_{i=1}^n \alpha^i \gamma^i, \quad r \in \mathbb{R}.$$

See Appendix A for proof.

Furthermore, we have $\eta_t \le \sum_{i=1}^n \alpha^i r^i$ and, therefore, Assumption 6 is satisfied. Therefore, we can bound ε_t and δ_t in Theorem 2 as

$$\varepsilon_t \le L^{\bar{\ell}} \bar{r} + 2\beta \quad \text{and} \quad \delta_t \le L^{\bar{f}} \bar{r} + 2\bar{\gamma}$$

where $\bar{r} = \sum_{i=1}^n \alpha^i r^i$ and $\bar{\gamma} = \sum_{i=1}^n \alpha^i \gamma^i$.

Therefore, we can upper bound the sub-optimality of using the certainty equivalent policy by the bound (15) in Theorem 2, which simplifies as follows:

$$W_t^{\mathcal{P},\mu^{\mathcal{E}}}(h_t) - W_t^{\mathcal{P}}(h_t) \le 2\Big[(T-t+1)(L^{\bar{c}}\bar{r} + 2\beta) + (L^{\bar{f}}\bar{r} + 2\bar{\gamma}) \sum_{\tau=t}^{T-1} \mathrm{Lip}(V_{\tau+1}^{\tilde{\mathcal{M}}})\Big]. \tag{33}$$

Remark 4 Such models can arise in situations where there is a local controller associated with each particle, and the local controller ensures that the $\|f^i\|_\infty \le \gamma^i$. Further, note that the bound in (33) depends on $\bar{r}$, which may be small even if some of the $\{r^i\}_{i=1}^n$ are large.

IV. Analysis

In this section we provide the analysis for our main result. As stated earlier, Theorem 1 is a special case of Theorem 2 obtained by taking $\tilde{\mathcal{S}} = \mathcal{S}$ and $\phi(s) = s$. Therefore, we only provide a proof of Theorem 2.

We start with some background on policy independent beliefs, and the AIS theory, followed by the key lemmas and proofs.

A. Policy independent beliefs

Consider an arbitrary history dependent policy μ for the model $\mathcal{P}$ defined in Sec. II. We define the following two *beliefs* which are commonly used in POMDPs:

- $b_{t|t}(\cdot|h_t)$ denotes the controller's posterior distribution on the current state S_t given the history h_t under the policy μ, i.e., for any Borel subset M_S of $\mathcal{S}$, $b_{t|t}(M_S|h_t) = \mathbb{P}^\mu(S_t \in M_S|h_t)$. The belief $b_{t|t}(\cdot|h_t)$ is referred to as the *belief state*. It is well known that it does not depend on the choice of the history dependent policy μ [2], [3].
- $b_{t+1|t}(\cdot, \cdot|h_t, a_t)$ denotes the controller's posterior distribution on the next state S_{t+1} and next observation Y_{t+1} given the history h_t and action a_t under policy μ. Note that for any Borel subsets M_S and M_Y of $\mathcal{S}$ and $\mathcal{Y}$, we have

$$b_{t+1|t}(M_S, M_Y|h_t, a_t) = \int_{\mathcal{S}} P_t(M_S, M_Y|s_t, a_t) b_{t|t}(ds_t|h_t).$$

  Since the belief state $b_{t|t}(\cdot|h_t)$ does not depend on the choice of the policy μ, it follows from the above relationship that the same holds for $b_{t+1|t}(\cdot, \cdot|h_t, a_t)$ as well. With a slight abuse of notation, we will continue to use $b_{t+1|t}$ to denote its marginals on $\mathcal{S}$ or $\mathcal{Y}$.

B. Approximate information states

The AIS theory [11] provides a framework to derive sub-optimality bounds for a class of approximate solutions to POMDPs. The key idea in this framework is the notion of an approximate information state, which we formally define below. Our definition is similar to that of [11] with two differences. First, the analysis in [11] was done under the assumption that the state and observation spaces are finite, while we are working with Borel spaces. So, we include a *measurable selection assumption* to ensure that the approximate dynamic program obtained from the AIS has a well-defined solution. Second, the analysis in [11] used general integral probability metrics (IPMs) [37]. We restrict our discussion to a specific choice of IPM (Wasserstein-1 distance), since that is the form that is used in our results.

The discussion below is for the general POMDP model $\mathcal{P}$ defined in Sec. II.

Definition 2 Given sequences $\varepsilon = (\varepsilon_1, \dots, \varepsilon_T)$ and $\delta = (\delta_1, \dots, \delta_{T-1}) \in \mathbb{R}^T_{\geq 0}$, a process $\{Z_t\}_{t\geq 1}$, $Z_t \in \mathcal{Z}$, is called an (ε, δ)-*approximate information state (AIS)* if there exist

- a sequence of history compression functions $\{\sigma_t^{\mathsf{AIS}}\}_{t=1}^T$, where $\sigma_t^{\mathsf{AIS}} \colon \mathcal{H}_t \to \mathcal{Z}$ with $Z_t = \sigma_t^{\mathsf{AIS}}(H_t)$
- a sequence of cost approximators $\{c_t^{\mathsf{AIS}}\}_{t=1}^T$, where $c_t^{\mathsf{AIS}} \colon \mathcal{Z} \times \mathcal{A} \to \mathbb{R}$
- a sequence of dynamics approximators $\{P_t^{\mathsf{AIS}}\}_{t=1}^{T-1}$, where $P_t^{\mathsf{AIS}} \colon \mathcal{Z} \times \mathcal{A} \to \Delta(\mathcal{Z})$

such that following three properties are satisfied:

(AP1) *Approximately sufficient for performance evaluation:* for any time $t \in \{1, \dots, T\}$ and any $h_t \in \mathcal{H}_t$ and $a_t \in \mathcal{A}$, we have

$$\big|\mathbb{E}[c_t(S_t, a_t)|h_t] - c_t^{\mathsf{AIS}}(\sigma_t^{\mathsf{AIS}}(h_t), a_t)\big| \leq \varepsilon_t$$

(AP2) *Approximately sufficient for predicting itself:* for any time $t \in \{1, \dots, T-1\}$ and any $h_t \in \mathcal{H}_t$ and $a_t \in \mathcal{A}$, define the stochastic kernel ν_t on $\mathcal{H}_t \times \mathcal{A}_t \to \Delta(\mathcal{Z})$ as follows: for any Borel measurable subset M_Z of $\mathcal{Z}$,

$$\begin{aligned}\nu_t(M_Z|h_t, a_t) &= \mathbb{P}(Z_{t+1} \in M_Z | h_t, a_t) \\ &= \int_{\mathcal{Y}} \mathbb{1}\{\sigma_{t+1}^{\mathsf{AIS}}(h_t, a_t, y_{t+1}) \in M_Z\} b_{t+1|t}(dy_{t+1}|h_t, a_t).\end{aligned}$$

Then, for any time $t \in \{1, \dots, T-1\}$, we have

$$d_{\mathrm{Was}}\big(\nu_t(\cdot|h_t, a_t), P_t^{\mathsf{AIS}}(\cdot|\sigma_t^{\mathsf{AIS}}(h_t), a_t)\big) \leq \delta_t.$$

(M) *Measurable selection:* The MDP $\mathcal{M}^{\mathsf{AIS}} := \langle \mathcal{Z}, \mathcal{A}, \{P_t^{\mathsf{AIS}}\}_{t=1}^{T-1}, \{c_t^{\mathsf{AIS}}\}_{t=1}^T, T\rangle$ satisfies measurable selection.

The tuple $(\sigma^{\mathsf{AIS}}, c^{\mathsf{AIS}}, P^{\mathsf{AIS}})$, where each component is a sequence, is called an *AIS-generator*.

We can write a dynamic program for $\mathcal{M}^{\mathsf{AIS}}$ where the value functions $\{V_t^{\mathsf{AIS}}\}_{t=1}^{T+1}$, $V_t^{\mathsf{AIS}} \colon \mathcal{Z} \to \mathbb{R}$, are defined as follows. We initialize $V_{T+1}^{\mathsf{AIS}}(z) = 0$ for all $z \in \mathcal{Z}$ and then recursively define for $t \in \{T, T-1, \dots, 1\}$

$$V_t^{\mathsf{AIS}}(z_t) = \min_{a\in\mathcal{A}} \left\{ c_t^{\mathsf{AIS}}(z_t, a) + \int_{\mathcal{Z}} V_{t+1}^{\mathsf{AIS}}(z') P_t^{\mathsf{AIS}}(dz'|z_t, a) \right\}. \tag{34}$$

The measurable selection condition (M) implies that there exists a measurable selector $\pi_t^{\mathsf{AIS}} \colon \mathcal{Z} \to \mathcal{A}$, $t \in \{1, \dots, T\}$, such that $\pi_t^{\mathsf{AIS}}(z_t)$ is an arg min of the right hand side of (34) and the functions V_t^{AIS} are measurable. From standard results in MDP theory [31], we know that the policy $\pi^{\mathsf{AIS}} = (\pi_1^{\mathsf{AIS}}, \dots, \pi_T^{\mathsf{AIS}})$ is an optimal policy for $\mathcal{M}^{\mathsf{AIS}}$.

The main result of the AIS theory is the following:

Theorem 3 *Define a history-dependent policy $\mu^{\mathsf{AIS}} = (\mu_1^{\mathsf{AIS}}, \dots, \mu_T^{\mathsf{AIS}})$ for the POMDP $\mathcal{P}$ as follows: for any $t \in \{1, \dots, T\}$ and any $h_t \in \mathcal{H}_t$, define*

$$\mu_t^{\mathsf{AIS}}(h_t) = \pi^{\mathsf{AIS}}(\sigma_t^{\mathsf{AIS}}(h_t)).$$

Then, for any $t \in \{1, \dots, T\}$ and $h_t \in \mathcal{H}_t$, we have

$$W_t^{\mathcal{P}, \mu^{\mathsf{AIS}}}(h_t) - W_t^{\mathcal{P}}(h_t) \leq 2\alpha_t \tag{35}$$

where

$$\alpha_t = \varepsilon_t + \sum_{\tau=t}^{T-1} \big[\delta_\tau \operatorname{Lip}(V_{\tau+1}^{\mathsf{AIS}}) + \varepsilon_{\tau+1}\big].$$

PROOF The result is the same as [11, Theorem 9], which was stated under the assumption that $\mathcal{S}$ and $\mathcal{A}$ are finite sets while we are working with Borel spaces. As argued earlier, the measurable selection condition ensures that V_t^{AIS} and π_t^{AIS} are well-defined and measurable. Under this assumption, the approximation bound follows from exactly the same analysis as in [11, Theorem 9]. ∎

C. Key lemmas

The main idea of our sub-optimality bounds is to show that the abstract state estimation functions $\mathcal{E}_t$, along with the per-step cost $\tilde{c}_t$ (defined in (10)) and dynamics $\tilde{P}_t$ (defined in (9)) of the abstract MDP $\tilde{\mathcal{M}}$ form an AIS generator for an appropriate choice of ε and δ. $(\mathcal{E}, \tilde{c}, \tilde{P})$.

We first show that $\mathcal{E}, \tilde{c}$ satisfy condition (AP1) of AIS.

Lemma 2 *Under Assumptions 5 and 6, for any $h_t \in \mathcal{H}_t$ and $a_t \in \mathcal{A}$, we have*

$$\big|\mathbb{E}[c_t(S_t, a_t)|h_t] - \tilde{c}_t(\mathcal{E}_t(h_t), a_t)\big| \leq F_t^c(\eta_t).$$

PROOF

$$\begin{aligned}&\big|\mathbb{E}[c_t(S_t, a_t)|h_t] - \tilde{c}_t(\mathcal{E}_t(h_t), a_t)\big| \\ &\qquad \leq \mathbb{E}\big[\big|c_t(S_t, a_t) - \tilde{c}_t(\mathcal{E}_t(h_t), a_t)\big| \mid h_t\big]\end{aligned} \tag{36}$$

We now consider the inner term for a fixed realization s_t,

$$\begin{aligned}&|c_t(s_t, a_t) - \tilde{c}_t(\mathcal{E}_t(h_t), a_t)| \\ &\overset{(a)}{=} \left| \int_{\phi^{-1}(\mathcal{E}_t(h_t))} \big[c_t(s_t, a_t) - c_t(s', a_t)\big] \lambda_t^c\big(ds'|\mathcal{E}_t(h_t)\big) \right| \\ &\overset{(b)}{\leq} \int_{\phi^{-1}(\mathcal{E}_t(h_t))} \big|c_t(s_t, a_t) - c_t(s', a_t)\big| \lambda_t^c\big(ds'|\mathcal{E}_t(h_t)\big) \\ &\overset{(c)}{\leq} \int_{\phi^{-1}(\mathcal{E}_t(h_t))} F_t^c(d_{\tilde{\mathcal{S}}}(\phi(s_t), \phi(s'))) \lambda_t^c\big(ds'|\mathcal{E}_t(h_t)\big) \\ &\overset{(d)}{=} F_t^c(d_{\tilde{\mathcal{S}}}(\phi(s_t), \mathcal{E}_t(h_t))).\end{aligned} \tag{37}$$

where (a) follows from definition of $\tilde{c}_t$, (b) follows from Jensen's inequality, (c) follows from Assumption 5 and (d) follows from the fact that for any $s' \in \phi^{-1}(\mathcal{E}_t(h_t))$, $\phi(s') = \mathcal{E}_t(h_t)$ and that $\lambda_t^c(\phi^{-1}(\mathcal{E}_t(h_t))|\mathcal{E}_t(h_t)) = 1$. Substituting (37) in (36), we get

$$\begin{aligned}&\big|\mathbb{E}[c_t(S_t, a_t)|h_t] - \tilde{c}_t(\mathcal{E}_t(h_t), a_t)\big| \\ &\qquad \leq \mathbb{E}[F_t^c(d_{\tilde{\mathcal{S}}}(\phi(S_t), \mathcal{E}_t(h_t)))|h_t] \\ &\qquad \overset{(e)}{\leq} F_t^c\big(\mathbb{E}[d_{\tilde{\mathcal{S}}}(\phi(S_t), \mathcal{E}_t(h_t)) \mid h_t]\big) \\ &\qquad \overset{(f)}{\leq} F_t^c(\eta_t)\end{aligned} \tag{38}$$

where (e) follows from Jensen's inequality and the concavity of F_t^c and (f) follows from the definition of η_t. ∎

In order to show that $\mathcal{E}, \tilde{P}$ satisfy condition (AP2) of AIS, we will use the following intermediate lemma.

Lemma 3 *Under Assumption 5, for any $s_t \in \mathcal{S}$, $\hat{s}_t \in \tilde{\mathcal{S}}$ and $a_t \in \mathcal{A}$, we have that*

$$d_{\mathrm{Was}}\big(P_{S,t}^{\phi}(\cdot|s_t, a_t), \tilde{P}_t(\cdot|\hat{s}_t, a_t)\big) \leq F_t^P\Big(d_{\tilde{\mathcal{S}}}\big(\phi(s_t), \hat{s}_t\big)\Big).$$

PROOF

$$
\begin{aligned}
&d_{\mathrm{Was}}\big(P^{\phi}_{S,t}(\cdot|s_t,a_t), \tilde{P}_t(\cdot|\hat{s}_t,a_t)\big)\\
&= d_{\mathrm{Was}}\big(P^{\phi}_{S,t}(\cdot|s_t,a_t), \int_{\phi^{-1}(\hat{s}_t)} P^{\phi}_{S,t}(\cdot|s',a_t)\lambda^P(ds'|\hat{s}_t)\big)\\
&\overset{(a)}{\le} \int_{\phi^{-1}(\hat{s}_t)} d_{\mathrm{Was}}\Big(P^{\phi}_{S,t}(\cdot|s_t,a_t), P^{\phi}_{S,t}(\cdot|s',a_t)\Big)\lambda^P(ds'|\hat{s}_t)\\
&\overset{(b)}{\le} \int_{\phi^{-1}(\hat{s}_t)} F^P_t\Big(d_{\tilde{\mathcal{S}}}\big(\phi(s_t),\phi(s')\big)\Big)\lambda^P(ds'|\hat{s}_t)\\
&= F^P_t\Big(d_{\tilde{\mathcal{S}}}\big(\phi(s_t),\hat{s}_t\big)\Big), \qquad (39)
\end{aligned}
$$

where (a) follows from the convexity of Wasserstein distance [38, Thm. 4.8] and (b) follows from Assumption 5. ∎

Next we define a stochastic kernel $\hat{\psi}_t\colon \mathcal{H}_t \times \mathcal{A}_t \to \Delta(\tilde{\mathcal{S}})$, which is analogous to ν_t defined in (AP2). For any $h_t \in \mathcal{H}_t$ and $a_t \in \mathcal{A}$ and Borel measurable subset $M_{\tilde{S}}$ of $\tilde{\mathcal{S}}$,

$$
\begin{aligned}
&\hat{\psi}_t(M_{\tilde{S}}|h_t,a_t) = \mathbb{P}(\mathcal{E}_{t+1}(H_{t+1}) \in M_{\tilde{S}}|h_t,a_t)\\
&= \int_{\mathcal{Y}} \mathbb{1}\{\mathcal{E}_{t+1}(h_t,a_t,y_{t+1}) \in M_{\tilde{S}}\} b_{t+1|t}(dy_{t+1}|h_t,a_t) \quad (40)
\end{aligned}
$$

which is the conditional probability distribution of $\hat{S}_{t+1} = \mathcal{E}_{t+1}(H_{t+1})$ given h_t, a_t.

We also define the stochastic kernel $\tilde{\psi}_t\colon \mathcal{H}_t \times \mathcal{A}_t \to \Delta(\tilde{\mathcal{S}})$, which is used in the proof of the next lemma. For any $h_t \in \mathcal{H}_t$ and $a_t \in \mathcal{A}$ and Borel measurable subset $M_{\tilde{S}}$ of $\tilde{\mathcal{S}}$,

$$
\begin{aligned}
&\tilde{\psi}_t(M_{\tilde{S}}|h_t,a_t) = \mathbb{P}(\tilde{S}_{t+1} \in M_{\tilde{S}}|h_t,a_t)\\
&= \int_{\mathcal{S}} \mathbb{1}\{\phi(s_{t+1}) \in M_{\tilde{S}}\} b_{t+1|t}(ds_{t+1}|h_t,a_t)\\
&= \int_{\mathcal{S}} P^{\phi}_{S,t}(M_{\tilde{S}}|s_t,a_t) b_{t|t}(ds_t|h_t,a_t), \qquad (41)
\end{aligned}
$$

which is the conditional probability distribution of $\phi(S_{t+1})$ given h_t, a_t.

The following lemma shows that $\mathcal{E}_t$, $\tilde{P}_t$ satisfy (AP2).

Lemma 4 *Under Assumptions 5 and 6, for any $h_t \in \mathcal{H}_t$ and $a_t \in \mathcal{A}$, we have*

$$
d_{\mathrm{Was}}(\hat{\psi}_t(\cdot|h_t,a_t), \tilde{P}_t(\cdot|\mathcal{E}_t(h_t),a_t)) \le F^P_t(\eta_t) + \eta_{t+1},
$$

where $\hat{\psi}_t(\cdot|h_t,a_t)$ is the probability distribution on $\tilde{\mathcal{S}}$ defined in (40).

PROOF Let $\hat{s}_t = \mathcal{E}_t(h_t)$. By triangle inequality, we have

$$
\begin{aligned}
&d_{\mathrm{Was}}(\hat{\psi}_t(\cdot|h_t,a_t), \tilde{P}_t(\cdot|\hat{s}_t,a_t))\\
&\quad \le d_{\mathrm{Was}}(\hat{\psi}_t(\cdot|h_t,a_t), \tilde{\psi}_t(\cdot|h_t,a_t))\\
&\qquad + d_{\mathrm{Was}}(\tilde{\psi}_t(\cdot|h_t,a_t), \tilde{P}_t(\cdot|\hat{s}_t,a_t)). \qquad (42)
\end{aligned}
$$

Now we consider the two terms separately. The first term of (42) can be bounded as follows.

$$
\begin{aligned}
&d_{\mathrm{Was}}(\hat{\psi}_t(\cdot|h_t,a_t), \tilde{\psi}_t(\cdot|h_t,a_t))\\
&\overset{(a)}{=} \inf_{(\hat{S},\tilde{S})\sim\Gamma(\hat{\psi}_t,\tilde{\psi}_t)} \mathbb{E}\big[d_{\tilde{\mathcal{S}}}(\hat{S},\tilde{S})\big]\\
&\overset{(b)}{\le} \mathbb{E}[d_{\tilde{\mathcal{S}}}\big(\mathcal{E}_{t+1}(H_{t+1}), \phi(S_{t+1})\big)|h_t,a_t]\\
&= \mathbb{E}[\mathbb{E}[d_{\tilde{\mathcal{S}}}\big(\mathcal{E}_{t+1}(H_{t+1}), \phi(S_{t+1})\big)|H_{t+1}]|h_t,a_t]\\
&\le \mathbb{E}[\eta_{t+1}|h_t,a_t] = \eta_{t+1} \qquad (43)
\end{aligned}
$$

where, in (a), $\Gamma(\hat{\psi}_t, \tilde{\psi}_t)$ denotes the set of all joint measures with marginals $\hat{\psi}_t(\cdot|h_t,a_t)$ and $\tilde{\psi}_t(\cdot|h_t,a_t)$, and in (b), we use the fact that conditioned on h_t, a_t, the marginal distributions of $\mathcal{E}_{t+1}(H_{t+1})$ and $\phi(S_{t+1})$ are $\hat{\psi}_t(\cdot|h_t,a_t)$ and $\tilde{\psi}_t(\cdot|h_t,a_t)$ respectively; therefore, the joint distribution on $(\mathcal{E}_{t+1}(H_{t+1}), \phi(S_{t+1}))$ conditioned on (h_t,a_t) lies in $\Gamma(\hat{\psi}_t, \tilde{\psi}_t)$.

The second term of (42) can be bounded as follows.

$$
\begin{aligned}
&d_{\mathrm{Was}}(\tilde{\psi}_t(\cdot|h_t,a_t), \tilde{P}_t(\cdot|\hat{s}_t,a_t))\\
&\quad = d_{\mathrm{Was}}\Big(\int_{\mathcal{S}} P^{\phi}_{S,t}(\cdot|s_t,a_t) b_{t|t}(ds_t|h_t), \tilde{P}_t(\cdot|\hat{s}_t,a_t)\Big)\\
&\quad \overset{(a)}{\le} \int_{\mathcal{S}} d_{\mathrm{Was}}\big(P^{\phi}_{S,t}(\cdot|s_t,a_t), \tilde{P}_t(\cdot|\hat{s}_t,a_t)\big) b_{t|t}(ds_t|h_t)\\
&\quad \overset{(b)}{\le} \int_{\mathcal{S}} F^P_t\big(d_{\tilde{\mathcal{S}}}(\phi(s_t),\hat{s}_t)\big) b_{t|t}(ds_t|h_t)\\
&\quad = \mathbb{E}[F^P_t\big(d_{\tilde{\mathcal{S}}}(\phi(S_t), \mathcal{E}_t(h_t))|h_t\big])\\
&\quad \overset{(c)}{\le} F^P_t\big(\mathbb{E}\big[d_{\tilde{\mathcal{S}}}(\phi(S_t), \mathcal{E}_t(h_t))|h_t\big]\big)\\
&\quad \le F^P_t(\eta_t). \qquad (44)
\end{aligned}
$$

where (a) follows from the convexity of Wasserstein distance [38, Thm. 4.8], and (b) follows from Lemma 3, and (c) follows from Jensen's inequality and the concavity of F^P_t. ∎

D. Proof of Theorem 2

Under Assumptions 5 and 6, Lemmas 2 and 4 ensure that conditions (AP1) and (AP2) of AIS are satisfied with $\varepsilon_t = F^c_t(\eta_t), \delta_t = F^P_t(\eta_t) + \eta_{t+1}$. Assumption 4 ensures that condition (M) of AIS is satisfied. Thus, the result follows from Theorem 3.

V. CONCLUSION

In this paper, we introduced a generalization of the certainty equivalence principle for control policies in partially observable Markov decision processes (POMDPs). Our approach applies optimal state-feedback policies from the fully observable MDP to state estimates, without restricting to specific types of estimators such as MMSE. We established theoretical performance bounds that characterize their degree of sub-optimality. Specifically, we leveraged the approximate information state (AIS) framework [11] to quantify the impact of estimation errors on control performance, deriving bounds in terms of the smoothness of the system dynamics and the per-step cost function.

To illustrate the practical relevance of our results, we examined several examples that demonstrate that certainty equivalent policies can perform near-optimally when state estimation errors are small. This suggests that in scenarios where exact optimal policies are computationally intractable, certainty equivalent policies offer a practical and efficient alternative, making effective use of available state estimates to achieve reliable decision-making while maintaining tractability.

REFERENCES

[1] B. Bozkurt, A. Mahajan, A. Nayyar, and Y. Ouyang, "Generalized certainty equivalence based policies in partially observable systems," in *IEEE Conference on Decision and Control*. IEEE, Dec. 2025.

[2] K. J. Åström, "Optimal control of Markov processes with incomplete state information," *Journal of Mathematical Analysis and Applications*, vol. 10, no. 1, pp. 174–205, Feb. 1965.

[3] R. D. Smallwood and E. J. Sondik, "The optimal control of partially observable Markov processes over a finite horizon," *Operations Research*, vol. 21, no. 5, pp. 1071–1088, Oct. 1973.

[4] C. H. Papadimitriou and J. N. Tsitsiklis, "The complexity of Markov decision processes," *Mathematics of operations research*, vol. 12, no. 3, pp. 441–450, 1987.

[5] G. Shani, J. Pineau, and R. Kaplow, "A survey of point-based pomdp solvers," *Autonomous Agents and Multi-Agent Systems*, vol. 27, no. 1, pp. 1–51, 2013.

[6] D. Burago, M. De Rougemont, and A. Slissenko, "On the complexity of partially observed Markov decision processes," *Theoretical Computer Science*, vol. 157, no. 2, pp. 161–183, 1996.

[7] C. Lusena, J. Goldsmith, and M. Mundhenk, "Nonapproximability results for partially observable Markov decision processes," *Journal of artificial intelligence research*, vol. 14, pp. 83–103, 2001.

[8] C. C. White III and W. T. Scherer, "Finite-memory suboptimal design for partially observed Markov decision processes," *Operations Research*, vol. 42, no. 3, pp. 439–455, 1994.

[9] B. Van Roy, Q. Dong, and L. Tang, "Simple agent, complex environment: Efficient reinforcement learning with agent states," *Journal of Machine Learning Research*, vol. 24, no. 170, pp. 1–54, 2023.

[10] A. Sinha and A. Mahajan, "Agent-state based policies in pomdps: Beyond belief-state mdps," in *Conference on Decision and Control*, Dec. 2024, pp. 6722–6735.

[11] J. Subramanian, A. Sinha, R. Seraj, and A. Mahajan, "Approximate information state for approximate planning and reinforcement learning in partially observed systems." *J. Mach. Learn. Res.*, vol. 23, no. 12, pp. 1–83, 2022.

[12] C. McDonald and S. Yüksel, "Robustness to incorrect priors and controlled filter stability in partially observed stochastic control," *SIAM J. Control Optim.*, vol. 60, no. 2, pp. 842–870, 2022.

[13] A. Kara and S. Yuksel, "Near optimality of finite memory feedback policies in partially observed markov decision processes," *Journal of Machine Learning Research*, vol. 23, no. 11, pp. 1–46, 2022.

[14] N. Golowich, A. Moitra, and D. Rohatgi, "Planning and learning in partially observable systems via filter stability," in *ACM Symposium on Theory of Computing*, ser. STOC 2023. New York, NY, USA: Association for Computing Machinery, 2023, p. 349–362.

[15] Q. Liu, A. Chung, C. Szepesvári, and C. Jin, "When is partially observable reinforcement learning not scary?" in *Conference on Learning Theory*. PMLR, 2022, pp. 5175–5220.

[16] W. Lee, N. Rong, and D. Hsu, "What makes some pomdp problems easy to approximate?" *Advances in neural information processing systems*, vol. 20, 2007.

[17] J. Guo, Z. Li, H. Wang, M. Wang, Z. Yang, and X. Zhang, "Provably efficient representation learning with tractable planning in low-rank pomdp," in *International Conference on Machine Learning*. PMLR, 2023, pp. 11 967–11 997.

[18] M. Belly, N. Fijalkow, H. Gimbert, F. Horn, G. A. Pérez, and P. Vandenhove, "Revelations: A decidable class of pomdps with omega-regular objectives," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 25, 2025, pp. 26 454–26 462.

[19] H. Theil, "Econometric models and welfare maximization," *Wirtschaftliches Archiv*, vol. 72, pp. 60–83, 1954.

[20] ——, "A note on certainty equivalence in dynamic planning," *Econometrica*, pp. 346–349, 1957.

[21] H. A. Simon, "Dynamic programming under uncertainty with a quadratic criterion function," *Econometrica: Journal of the Econometric Society*, pp. 74–81, 1956.

[22] Y. Bar-Shalom and E. Tse, "Dual effect, certainty equivalence, and separation in stochastic control," *IEEE Transactions on Automatic Control*, vol. 19, no. 5, pp. 494–500, 1974.

[23] M. S. Derpich and S. Yüksel, "Dual effect, certainty equivalence, and separation revisited: A counterexample and a relaxed characterization for optimality," *IEEE Trans. Autom. Control*, vol. 68, no. 2, pp. 1259–1266, 2022.

[24] Y. Li and D. Bertsekas, "Semilinear dynamic programming: Analysis, algorithms, and certainty equivalence properties," *arXiv preprint arXiv:2501.04668*, 2025.

[25] P. Whittle, "The risk-sensitive certainty equivalence principle," *Journal of applied probability*, vol. 23, no. A, pp. 383–388, 1986.

[26] M. Hardt and B. Recht, *Patterns, predictions, and actions: Foundations of machine learning*. Princeton University Press, 2022.

[27] D. Bertsekas, *Dynamic programming and optimal control: Volume I*. Athena scientific, 2012.

[28] B. Wolfe, M. R. James, and S. Singh, "Approximate predictive state representations," in *Int. Conf. Auton. Agents Multiagent Syst.*, 2008, pp. 363–370.

[29] W. Hamilton, M. M. Fard, and J. Pineau, "Efficient learning and planning with compressed predictive states," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3395–3439, 2014.

[30] P. S. Castro, P. Panangaden, and D. Precup, "Equivalence relations in fully and partially observable Markov decision processes," in *Int. Jt. Conf. Artif. Intell.*, 2009, pp. 1653–1658.

[31] O. Hernández-Lerma and J. B. Lasserre, *Discrete-time Markov control processes: basic optimality criteria*. Springer Science & Business Media, 2012.

[32] K. Hinderer, "Lipschitz continuity of value functions in Markovian decision processes," *Mathematical Methods of Operations Research*, vol. 62, no. 1, pp. 3–22, 2005.

[33] C. Gelada, S. Kumar, J. Buckman, O. Nachum, and M. G. Bellemare, "Deepmdp: Learning continuous latent space models for representation learning," in *ICML*. PMLR, 2019, pp. 2170–2179.

[34] A. Molin and S. Hirche, "Suboptimal event-triggered control for networked control systems," pp. 277–289, 2014.

[35] M. Mazo and P. Tabuada, "Decentralized event-triggered control over wireless sensor/actuator networks," *IEEE Transactions on Automatic Control*, vol. 56, no. 10, pp. 2456–2461, 2011.

[36] W. P. Heemels, K. H. Johansson, and P. Tabuada, "An introduction to event-triggered and self-triggered control," in *IEEE Conference on Decision and Control*. IEEE, Dec. 2012, pp. 3270–3285.

[37] A. Müller, "How does the value function of a Markov decision process depend on the transition probabilities?" *Math. Oper. Res.*, vol. 22, no. 4, pp. 872–885, 1997.

[38] C. Villani, *Optimal transport: old and new*. Springer, 2009.

APPENDIX A
PROOF OF LEMMA 1

We prove the two parts separately.

1) Proof of (13): Arbitrarily pick $s, s' \in \mathcal{S}$. Let $m = \phi(s), m' = \phi(s') \in \tilde{\mathcal{S}} = \mathbb{R}$. Then, by triangle inequality we have

$$\begin{aligned}|c(s,a) - c(s',a)| &\le |\bar{\ell}(m,a) - \bar{\ell}(m',a)| + |\ell(s,a) - \ell(s',a)| \\ &\le L^{\bar{\ell}}|m - m'| + 2\beta := F_t^c(|m - m'|).\end{aligned}$$

2) Proof of (12): Arbitrarily pick $(m,a) \in \mathcal{S} \times \mathcal{A}$. Let $M(w) = \bar{f}(m,a,w) + \sum_{i=1}^n \alpha^i f^i(m\mathbf{1}_n, a, w)$ and $M'(w') = \bar{f}(m',a,w') + \sum_{=1}^n \alpha^i \bar{f}^i(m'\mathbf{1}_n, a, w')$. The left hand side of (12) is the Wasserstein distance between random variables $M(W)$ and $M'(W')$, where W and W' are identically distributed random variables with marginal distribution P_W. Let Γ denote all joint couplings between W and W' such that the marginals are P_W. Then,

$$\begin{aligned}d_{\mathrm{Was}}(M(W), M'(W')) &= \inf_{(W,W')\sim\Gamma} \mathbb{E}[|M(W) - M(W')|] \\ &\le \mathbb{E}[|M(W) - M'(W)|]\end{aligned}$$

where in the last equation we have chosen a specific coupling $W = W'$.

Now observe that

$$\begin{aligned}
&\mathbb{E}[|M(W) - M'(W)|] \\
&= \mathbb{E}\Bigg[\Bigg|\bar{f}(m, a, W) - \bar{f}(m', a, W) \\
&\quad + \sum_{i=1}^{n} \alpha^i \big(f^i(m\mathbf{1}_n, a, W) - f^i(m'\mathbf{1}_n, a, W)\big)\Bigg|\Bigg] \\
&\le \mathbb{E}\big[|\bar{f}(m, a, W) - \bar{f}(m', a, W)|\big] \\
&+ \sum_{i=1}^{n} \alpha^i \mathbb{E}\Bigg[\Big|\big(f^i(m\mathbf{1}_n, a, W) - f^i(m'\mathbf{1}_n, a, W)\Big|\Bigg] \\
&\le L^{\bar{f}}|m - m'| + 2\sum_{i=1}^{n} \alpha^i \gamma^i := F_t^P(|m - m'|)
\end{aligned}$$

where the last inequality holds from Assumption 7.

Berk Bozkurt (Student Member, IEEE) received his BSc degree in Electrical and Electronics Engineering from Bilkent University, Ankara, Turkey, in 2021 and and the M.Sc. degree from the Electrical and Computer Engineering Department, McGill University, Montreal, Quebec, Canada. His research interests include reinforcement learning, game theory, stochastic control and Markov decision theory.

Aditya Mahajan (Senior Member, IEEE) is Professor of Electrical and Computer Engineering at McGill University, Montreal, Canada. He received the B.Tech degree in Electrical Engineering from the Indian Institute of Technology, Kanpur, India in 2003 and the MS and PhD degrees in Electrical Engineering and Computer Science from the University of Michigan, Ann Arbor, USA in 2006 and 2008, respectively. He serves or has served as Associate Editor of Transactions on Automatic Control, Control Systems Letters, and Math. of Control, Signal, and Systems. His research interests include learning and control of partially observable and multi-agent systems.

Ashutosh Nayyar (Senior Member, IEEE) is an Associate Professor of Electrical and Computer Engineering at the University of Southern California. He received a M.S. degree in electrical engineering and computer science, a M.S. degree in applied mathematics, and a Ph.D. degree in electrical engineering and computer science from the University of Michigan, Ann Arbor, MI, USA, in 2008, 2011, and 2011, respectively. His research interests are in decentralized stochastic control, decentralized decision-making in sensing and communication systems, reinforcement learning, game theory and mechanism design.

Yi Ouyang received the B.S. degree in Electrical Engineering from the National Taiwan University, Taipei, Taiwan in 2009, and the M.Sc and Ph.D. in Electrical Engineering and Computer Science at the University of Michigan, in 2012 and 2015, respectively. He is currently a researcher at Preferred Networks, Burlingame, CA. His research interests include reinforcement learning, stochastic control, and stochastic dynamic games.