Generalized certainty equivalence based policies in partially observable 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 and allow for any state estimate rather than restricting to MMSE estimates. In such settings, the certainty equivalent policy is not optimal. For models with Lipschitz cost and dynamics, we derive upper bounds on the sub-optimality of certainty equivalent policies in terms of expected error of the proposed estimator. We present several examples to illustrate the results.**

I. Introduction

Consider the optimal output feedback control of a discrete-time linear system driven by an independent noise process and incurring a quadratic per-step cost. The model is sometimes also called the LQG optimal control problem but we are not imposing the Gaussian assumption on the noise. As is well known, the optimal control policy for this problem has the following structure: the optimal action at each time is a linear function of the 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 assuming that the realization of all future random variables equals to their conditional mean (which, under the assumption that the noise process is independent across time, is equivalent to assuming that the realization of all future random variables equals to their means). This result is typically called the *certainty equivalence principle of stochastic control* [1]–[5].

In this paper, we present a generalization of the certainty equivalence principle to general partially observable Markov decision processes (POMDPs) [6], [7]. Our generalization is based on a slightly different interpretation of the certainty equivalence principle where we view the optimal output feedback control policy as follows: the optimal action at each time is obtained by evaluating the optimal *state-feedback* policy of the *stochastic* system (obtained by assuming that the decision maker perfectly observes the state of the system) at the MMSE (minimum mean square error) estimate of the state. 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 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 observations and actions at the decision maker until time t, let $\mathcal{E}_t(h_t)$ denote the MMSE estimator of the state given the history h_t. Then, the standard result for LQG optimal control is that

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

As mentioned earlier, the model $\mathcal{M}$ is stochastic rather than deterministic. But we will still call the above result as the certainty equivalence principle. Similar views on the certainty equivalence principle have been used in the reinforcement learning and adaptive control literature [8].

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

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

In this general setting, we define the generalized certainty equivalence policy $\mu^{\mathsf{GCE}} = (\mu_1^{\mathsf{GCE}}, \dots, \mu_T^{\mathsf{GCE}})$ as

$$\mu_t^{\mathsf{GCE}}(h_t) = \pi_t^{\mathcal{M}}(\mathcal{E}_t(h_t)). \tag{1}$$

Clearly, in general, μ^{GCE} is not optimal. Our main result is to characterize the degree of sub-optimality of the generalized certainty equivalence policy μ^{GCE} and illustrate via examples that such a policy is an attractive approximation policy in some situations.

Our results may be viewed as an instance of characterizing the sub-optimality gap of approximate policies for POMDPs. There is a rich literature on deriving such sup-optimality gaps such as using tools from predictive state representation [9],

Berk Bozkurt is with INLAN, Montreal, QC, Canada. (email: berk.bozkurt@mail.mcgill.ca). Aditya Mahajan is with the Dept. of Electrical and Computer Engineering, McGill University, Montreal, Canada. (email: aditya.mahajan@mcgill.ca). Ashutosh Nayyar is with the Dept. of Electrical and Computer Engineering, USC, Los Angeles, USA. (email: ashutosn@usc.edu). Yi Ouyang is with Atmanity, Santa Clara, CA, USA. (email: ouyangyii@gmail.com). The work of AM was supported in part by NSERC under Grant RGPIN-2021-0351. The work of AN was supported in part by NSF under Grant ECCS 2025732 and Grant ECCS 1750041.

[10], bisimulation metrics [11], approximation information states (AIS) [12], and filter stability [13], [14]. Our analysis is based on AIS-based approximation bounds of [12].

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 short hand for the sequence $(S_1, \dots, S_t)$. $\mathbb{P}(\cdot)$ denotes the probability of an event and $\mathbb{E}[\cdot]$ denotes the expectation of a random variable. We use the notation of the form $\mathbb{P}(S_{t+1} \in M_S | s_t, a_t)$ as a short hand for $\mathbb{P}(S_{t+1} \in M_S | S_t = s_t, A_t = a_t)$. Similar notation is used for conditional expectations. The Wasserstein-1 distance on the space of probability measures is denoted by d_{Was}.

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 all bounded and measurable real-valued functions on $\mathcal{X}$. For square symmetric matrices P and Q, $\mathsf{P} \preceq \mathsf{Q}$ means that $\mathsf{Q} - \mathsf{P}$ is positive semi-definite. The Lipschitz constant of a function f is denoted by $\mathrm{Lip}(f)$.

II. A motivating example

Consider a linear system that starts from a known initial state s_1 and, for $t > 1$, has the dynamics

$$s_{t+1} = \mathsf{A} s_t + \mathsf{B} a_t + w_t, \quad y_t = s_t + n_t$$

where $s_t \in \mathbb{R}^{d_s}$, $a_t \in \mathbb{R}^{d_a}$, $\mathsf{A} \in \mathbb{R}^{d_s \times d_s}$, $\mathsf{B} \in \mathbb{R}^{d_s \times d_a}$, and $\{w_t\}_{t=1}^{T-1}$ and $\{n_t\}_{t=1}^{T}$ are processes that are independent across time and also mutually independent. We assume that the noises are zero mean and have covariances W_t and N_t (but we do assume that they have a Gaussian density). The objective is to find a policy $\mu = (\mu_1, \dots, \mu_{T-1})$, where $\mu_t \colon h_t \mapsto a_t$ to minimize the expected total cost

$$J(\mu) \coloneqq \mathbb{E}^{\mu}\left[\sum_{t=1}^{T-1} \left[s_t^{\mathsf{T}} \mathsf{Q}_t s_t + a_t^{\mathsf{T}} \mathsf{R}_t a_t\right] + s_T \mathsf{Q}_T^{\mathsf{T}} s_T\right]$$

where $\mathsf{Q}_t \in \mathbb{R}^{d_s \times d_s}$ are symmetric and positive semi-definite matrices and $\mathsf{R}_t \in \mathbb{R}^{d_a \times d_a}$ are symmetric and positive definite matrices. T

The above model is a special case of the standard LQG optimal control problem and therefore the optimal policy $\mu^{\mathcal{P}}$ is given as follows. Let $\{\mathsf{P}_t\}_{t \geq 1}$ be the solution of an appropriately defined Riccati equation and $\mathsf{K}_t = (\mathsf{R}_t + \mathsf{B}^{\mathsf{T}} \mathsf{P}_{t+1} \mathsf{B})^{-1} \mathsf{B}^{\mathsf{T}} \mathsf{P}_{t+1} \mathsf{A}$. Then, the optimal policy is given by

$$\mu_t^{\mathcal{P}}(h_t) = -\mathsf{K}_t \hat{s}_t$$

where $\hat{s}_t$ is the MMSE state estimate computed via non-linear filtering (or via Kalman filtering when the noise processes are Gaussian). Furthermore, the performance of the optimal policy is given by

$$J(\mu^{\mathcal{P}}) = \sum_{t=1}^{T-1} \left[\mathrm{Tr}(\mathsf{W}_t \mathsf{P}_{t+1}) + \mathrm{Tr}(\Sigma_t \mathsf{K}_t^{\mathsf{T}} (\mathsf{R}_t + \mathsf{B}^{\mathsf{T}} \mathsf{P}_{t+1} \mathsf{B}) \mathsf{K}_t)\right]$$

where Σ_t is the covariance of $(s_t - \hat{s}_t)$.

In general, implementing the optimal policy requires computing the MMSE estimates $\hat{s}_t$ via non-linear filtering, which can be computationally challenging (in the case when the noises are non-Gaussian). An alternative and easy-to-implement policy is a generalized certainty equivalence policy which uses the last observation as a state estimate, i.e., $\mathcal{E}_t(h_t) = y_t$ and $\mu_t^{\mathsf{GCE}}(h_t) = -\mathsf{K}_t y_t$ with the same feedback gain K_t as before. Simple algebra shows that the performance of this generalized certainty equivalent policy is

$$J(\mu^{\mathsf{GCE}}) = \sum_{t=1}^{T-1} \left[\mathrm{Tr}(\mathsf{W}_t \mathsf{P}_{t+1}) + \mathrm{Tr}(\mathsf{N}_t \mathsf{K}_t^{\mathsf{T}} (\mathsf{R}_t + \mathsf{B}^{\mathsf{T}} \mathsf{P}_{t+1} \mathsf{B}) \mathsf{K}_t)\right].$$

Thus, we obtain that the sub-optimality gap of μ^{GCE} is

$$J(\mu^{\mathsf{GCE}}) - J(\mu^{\mathcal{P}}) = \sum_{t=1}^{T-1} \mathrm{Tr}((\mathsf{N}_t - \Sigma_t) \mathsf{K}_t^{\mathsf{T}} (\mathsf{R}_t + \mathsf{B}^{\mathsf{T}} \mathsf{P}_{t+1} \mathsf{B}) \mathsf{K}_t),$$

which is small when the observation noise has a small covariance.

III. System model

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 system, $Y_t \in \mathcal{Y}$ denote the observation of controller, and $A_t \in \mathcal{A}$ denote the control action taken by the controller at time t. We assume that $\mathcal{S}$ is a metric space with a metric $d_{\mathcal{S}}$.

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 $c_{\max}$ such that $\sup_{s \in \mathcal{S}, a \in \mathcal{A}} |c_t(s, a)| \leq c_{\max}$.

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 realization 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 $W_t^{\mathcal{P},\mu}(h_t) = \mathbb{E}^{\mu}[\sum_{\tau=t}^{T} c_\tau(s_\tau, a_\tau) | h_t]$ where $\mathbb{E}^{\mu}$ denotes expectation with respect to a joint 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 solve a POMDP is to use a belief-state based dynamic program [6], [7]. As discussed

in the introduction, we are interested in obtaining approximately optimal policies that are motivated by the certainty equivalence principle. We explain the approximation approach in the next section.

A. Generalized 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.

Definition 1 (Measurable selection) An MDP with state and action spaces $\mathcal{S}$ and $\mathcal{A}$, per-step cost $\{c_t\}_{t=1}^T$ and dynamics $\{P_{S,t}\}_{t=1}^{T-1}$ 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

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

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 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 [15].

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 MMSE (minimum mean square error) 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. Note that these estimates need not be "good", just that they should generate an estimate that belongs to the state space.

We call a history-dependent policy $\mu^{\mathcal{E}} = (\mu_1^{\mathcal{E}}, \dots, \mu_T^{\mathcal{E}})$ *generalized certainty equivalent* with respect to $\{\mathcal{E}_t\}_{t\geq 1}$ if

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

As argued earlier, such policies are optimal in the LQ setting but are, in general, not optimal. We are interested in the following problem.

Problem 1 *Characterize the sub-optimality gap of generalized certainty equivalent policies for POMDPs, i.e., bound*

$$\left\| W_t^{\mathcal{P}} - W_t^{\mathcal{P},\mu^{\mathcal{E}}} \right\|_\infty = \sup_{h_t\in\mathcal{H}_t} \left| W_t^{\mathcal{P}}(h_t) - W_t^{\mathcal{P},\mu^{\mathcal{E}}}(h_t) \right|.$$

B. Some examples

As seen from the motivating example in Sec. II, certainty equivalent policies are attractive when there is "small observation noise". In general, the choice of a good estimator depends on the observation model. We present some examples to illustrate observation models (and corresponding estimators) where generalized certainty equivalent policies may be useful.

Example 1 Consider a POMDP where $\mathcal{Y} = \mathcal{S}$ and the observation model is such that $d_{\mathcal{S}}(Y_t, S_t) \leq r, \quad$ a.s. Suppose the state estimate is simply the last observation, i.e., $\mathcal{E}_t(h_t) = y_t$. Then, the generalized certainty equivalent policy is $\mu_t^{\mathcal{E}}(h_t) = \pi_t^{\mathcal{M}}(y_t)$.

Example 2 Consider a POMDP where $\mathcal{Y} = \mathcal{S}$ and the observation model is such that the controller perfectly observes the state S_t with probability $(1-p)$ and gets a random observation with probability p, i.e., $Y_t = \begin{cases} S_t, & \text{w.p. } (1-p) \\ U_t, & \text{w.p. } p \end{cases}$ where $\{U_t\}_{t\geq 1}$ is a sequence of independent and identically distributed random variables uniformly distributed on $\mathcal{S}$. Similar to Example 1, if the state estimate is chosen as the last observation, i.e., $\mathcal{E}_t(h_t) = y_t$, then the generalized certainty equivalent policy is $\mu_t^{\mathcal{E}}(h_t) = \pi_t^{\mathcal{M}}(y_t)$.

Example 3 Consider the MDP learning setting where the underlying system with state $\tilde{S}_t \in \tilde{\mathcal{S}}$ depends on an unknown parameter $\theta \in \Theta$ with the state transition kernel $P_{\tilde{\mathcal{S}},t}(\cdot|\tilde{s}_t, a_t; \theta)$, the cost function $c_t(\tilde{s}_t, a_t; \theta)$, and the optimal state-feedback policy $\pi_t^{\mathcal{M}}(\tilde{s}_t; \theta)$, all parametrized by θ, which is not directly observed. This parameter can be estimated/learned using various estimators, such as the MMSE estimator $\hat{\theta}_t = \mathbb{E}[\theta|h_t]$. By setting the overall system state $S_t = (\tilde{S}_t, \theta)$, the MDP learning problem becomes a POMDP with observation $Y_t = \tilde{S}_t$. Then, the state estimate corresponding to the MMSE estimator is given by $\mathcal{E}_t(h_t) = (\tilde{s}_t, \hat{\theta}_t)$, and the generalized certainty equivalent policy is $\mu_t^{\mathcal{E}}(h_t) = \pi_t^{\mathcal{M}}(\tilde{s}_t; \hat{\theta}_t)$.

Example 4 Consider a first-order linear system, i.e., a system with $\mathcal{S} = \mathcal{Y} = \mathcal{A} = \mathbb{R}$, where the dynamics are deterministic and given by $S_{t+1} = S_t + A_t$, and the observations are given by $Y_t = S_t + N_t$, where $\{N_t\}_{t\geq 1}$ is an independent and identically distributed process. We assume that the metric on the state space is $d_{\mathcal{S}}(s_1, s_2) = |s_1 - s_2|$ and $|N_t| \leq r$. The cost function is general, and not-necessarily quadratic.

Suppose the estimator is chosen as

$$\mathcal{E}(h_t) = \frac{1}{t}\left[\sum_{\tau=1}^{t}(y_\tau) - \sum_{\tau=1}^{t-1}(t-\tau)a_\tau\right] + \sum_{\tau=1}^{t-1} a_\tau.$$

Then, the generalized certainty equivalent policy is $\mu_t^{\mathcal{E}}(h_t) = \pi_t^{\mathcal{M}}(\mathcal{E}(h_t))$.

IV. Background

In this section we cover some of the background material needed to present our main result. We start by a discussion of some policy independent beliefs that are used in the analysis of POMDPs. We then introduce integral probability metrics (IPMs) [16], and provide a brief overview of the AIS

theory [12], which is the framework that we use to derive our sub-optimality bounds.

A. Policy independent beliefs

Consider an arbitrary history dependent policy μ for the model $\mathcal{P}$ defined in Sec. III. 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 know that it does not depend on the choice of the history dependent policy μ [6], [7].
- $b_{t+1|t}(\cdot,\cdot|h_t, a)$ 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 $\mathcal{Y}$,

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

Since the belief state $b_{t|t}(\cdot|h_t)$ does not depend on the choice of the policy μ, the above relationship implies that neither does $b_{t+1|t}(\cdot,\cdot|h_t, a_t)$. 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 [12] 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 [12] with one difference. The analysis in [12] was done under the assumption that the state and observation spaces are finite valued, while we are in general state spaces. So, we include a *measurable selection assumption* to ensure that the approximate dynamic program obtained from the AIS has a well defined solution.

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

Definition 2 Let $(\mathcal{Z}, d_{\mathcal{Z}})$ be a metric space. Given $\varepsilon = (\varepsilon_1, \dots, \varepsilon_T), \delta = (\delta_1, \dots, \delta_{T-1}) \in \mathbb{R}^T_{\geq 0}$, 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 $\{\sigma^{\mathsf{AIS}}_t\}_{t=1}^T$, where $\sigma^{\mathsf{AIS}}_t \colon \mathcal{H}_t \to \mathcal{Z}$ with $Z_t = \sigma^{\mathsf{AIS}}_t(H_t)$
- a sequence of cost approximators $\{c^{\mathsf{AIS}}_t\}_{t=1}^T$, where $c^{\mathsf{AIS}}_t \colon \mathcal{Z} \times \mathcal{A} \to \mathbb{R}$
- a sequence of dynamics approximators $\{P^{\mathsf{AIS}}_t\}_{t=1}^{T-1}$, where $P^{\mathsf{AIS}}_t \colon \mathcal{Z} \times \mathcal{A} \to \Delta(\mathcal{Z})$

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

$$\left|\mathbb{E}[c_t(S_t, a_t)|h_t] - c^{\mathsf{AIS}}_t(\sigma^{\mathsf{AIS}}_t(h_t), a_t)\right| \leq \varepsilon_t$$

(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 $\mathcal{Z}$,

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

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

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

(M) *Measurable selection:* The MDP model with state space $\mathcal{Z}$, action space $\mathcal{A}$, per-step costs $\{c^{\mathsf{AIS}}_t\}_{t=1}^T$ and dynamics $\{P^{\mathsf{AIS}}_t\}_{t=1}^T$ 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*.

Observe that components $(c^{\mathsf{AIS}}, P^{\mathsf{AIS}})$ of an AIS-generator define a model $\mathcal{M}^{\mathsf{AIS}}$ for an MDP. We can therefore write a dynamic program for this model where the value function $\{V^{\mathsf{AIS}}_t\}_{t=1}^{T+1}$, $V^{\mathsf{AIS}}_t \colon \mathcal{Z} \to \mathbb{R}$, are defined as follows. We initialize $V^{\mathsf{AIS}}_{T+1}(z) = 0$ for all $z \in \mathcal{Z}$ and then recursively define for $t \in \{T, T-1, \dots, 1\}$

$$V^{\mathsf{AIS}}_t(z_t) = \min_{a \in \mathcal{A}} \left\{ c^{\mathsf{AIS}}_t(z_t, a) + \int_{\mathcal{Z}} P^{\mathsf{AIS}}_t(dz'|z_t, a) V^{\mathsf{AIS}}_{t+1}(z') \right\}. \tag{4}$$

The measurable selection condition (M) implies that there exists a measurable selector $\pi^{\mathsf{AIS}}_t \colon \mathcal{Z} \to \mathcal{A}$, $t \in \{1, \dots, T\}$, such that $\pi^{\mathsf{AIS}}_t(z_t)$ is an arg min of the right hand side of (4) and the functions V^{AIS}_t are measurable. From standard results in MDP theory [15], we know that the policy $\pi^{\mathsf{AIS}} = (\pi^{\mathsf{AIS}}_1, \dots, \pi^{\mathsf{AIS}}_T)$ is an optimal policy for $\mathcal{M}^{\mathsf{AIS}}$.

The main result of the AIS theory is the following:

Theorem 1 *Define a history-dependent policy $\mu^{\mathsf{AIS}} = (\mu^{\mathsf{AIS}}_1, \dots, \mu^{\mathsf{AIS}}_T)$ for the POMDP $\mathcal{P}$ as follows: for any $t \in \{1, \dots, T\}$ and any $h_t \in \mathcal{H}_t$, define*

$$\mu^{\mathsf{AIS}}_t(h_t) = \pi^{\mathsf{AIS}}(\sigma^{\mathsf{AIS}}_t(h_t)).$$

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

$$W^{\mathcal{P}, \mu^{\mathsf{AIS}}}_t(h_t) - W^{\mathcal{P}}_t(h_t) \leq 2\alpha_t \tag{5}$$

where

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

PROOF (SKETCH) As argued earlier, the measurable selection condition ensures that V^{AIS}_t and π^{AIS}_t are well-defined and measurable. The approximation bound follows from exactly the same analysis as in [12, Theorem 9]. ∎

V. APPROXIMATION BOUNDS

The main idea of our sub-optimality bounds for Problem 1 is to show that the process $\{\hat{S}_t\}_{t\geq 1}$, where $\hat{S}_t = \mathcal{E}_t(H_t)$ is an (ε, δ) AIS for appropriate choice of ε and δ. Recall that $d_{\mathcal{S}}$ is the metric on the state space $\mathcal{S}$.

We impose the following assumption on the model.

Assumption 2 There exist finite constants $\{L_t^c\}_{t=1}^T$ and $\{L_t^P\}_{t=1}^T$ such that for any $s_1, s_2 \in \mathcal{S}$ and $a \in \mathcal{A}$, we have

$$\big|c_t(s_1,a) - c_t(s_2,a)\big| \le L_t^c d_{\mathcal{S}}(s_1,s_2), \tag{6}$$

$$d_{\mathrm{Was}}(P_{S,t}(\cdot|s_1,a), P_{S,t}(\cdot|s_2,a)) \le L_t^P d_{\mathcal{S}}(s_1,s_2). \tag{7}$$

These are standard assumptions for smoothness of the per-step cost and the dynamics, and imply smoothness (Lipschitz continuity) of the value function of model $\mathcal{M}$ [17].

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

$$\eta_t := \sup_{h_t \in \mathcal{H}_t} \mathbb{E}[d_{\mathcal{S}}(S_t, \mathcal{E}_t(h_t))|h_t]. \tag{8}$$

We state the following lemma (proofs are omitted due to space limitations).

Lemma 1 *Under Assumption 2, for any $h_t \in \mathcal{H}_t$ and $a_t \in \mathcal{A}$, we have*

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

Given a $h_t \in \mathcal{H}_t$ and $a_t \in \mathcal{A}$, define the stochastic kernel $\hat{\nu}_t : \mathcal{H}_t \times \mathcal{A}_t \to \Delta(\mathcal{S})$ as follows: for any Borel measurable subset M_S of $\mathcal{S}$,

$$\hat{\nu}_t(M_S|h_t,a_t) = \mathbb{P}(\mathcal{E}_{t+1}(H_{t+1}) \in M_S | h_t, a_t) \tag{9}$$

$$= \int_{\mathcal{Y}} \mathbb{1}\{\mathcal{E}_{t+1}(h_t,a_t,y_{t+1}) \in M_S\} b_{t+1|t}(dy_{t+1}|h_t,a_t). \tag{10}$$

The interpretation of $\hat{\nu}_t(\cdot|h_t,a_t)$ is that it is the conditional probability distribution of $\hat{S}_{t+1} = \mathcal{E}_{t+1}(H_{t+1})$ given h_t, a_t.

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

$$d_{\mathrm{Was}}(\hat{\nu}_t(\cdot|h_t,a_t), P_{S,t}(\cdot|\mathcal{E}_t(h_t),a_t)) \le L_t^P \eta_t + \eta_{t+1},$$

where $\hat{\nu}_t(\cdot|h_t,a_t)$ is the probability distribution on $\mathcal{S}$ defined in (10) and $P_{S,t}(\cdot|\mathcal{E}_t(h_t),a_t)$ is the distribution of the next state (i.e. S_{t+1}) if the current state is $\mathcal{E}_t(h_t)$ and the current action is a_t.

Theorem 2 *Under Assumptions 1 and 2, $(\mathcal{E}, c, P)$ is an (ε,δ)-AIS-generator with*

$$\varepsilon_t = L_t^c \eta_t, \quad \delta_t = L_t^P \eta_t + \eta_{t+1}.$$

Consequently, we have that for the generalized certainty equivalent policy $\mu^{\mathcal{E}}$ (defined in (3)),

$$W_t^{\mathcal{P},\mu^{\mathcal{E}}}(h_t) - W_t^{\mathcal{P}}(h_t) \le 2\alpha_t \tag{11}$$

where

$$\alpha_t = \varepsilon_t + \sum_{\tau=t}^{T-1} \big[\delta_t \operatorname{Lip}(V_{t+1}^{\mathcal{M}}) + \varepsilon_{\tau+1}\big] \tag{12}$$

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

PROOF Under Assumption 2, Lemmas 1 and 2 ensure that conditions (AP1) and (AP2) of AIS are satisfied with $\varepsilon_t = L_t^c \eta_t, \delta_t = L_t^P \eta_t + \eta_{t+1}$. Assumption 1 ensures that condition (M) of AIS is satisfied. Thus, the result follows from Theorem 1. ∎

Remark 1 Following the argument in [17], it can be shown that under Assumptions 1 and 2, the optimal value function $V_t^{\mathcal{M}}$ for the MDP $\mathcal{M}$ is Lipschitz with constant $\operatorname{Lip}(V_t^{\mathcal{M}}) \le L_c(1 + L_p + L_p^2 + \cdots + L_p^{T-t})$. Using this inequality in (12), gives an upper bound on α_t.

VI. Some illustrative examples

In this section we apply our results to the examples presented in Section III-B to derive explicit bounds on the sub-optimality of generalized certainty equivalent policies for specific observation models.

A. Example 1

For the observation model given in Example 1,

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

Hence, we have $\eta_t \le r$. Thus, if the model $\mathcal{M}$ satisfies Assumptions 1 and 2, then $(\mathcal{E}, c, P)$ is an AIS generator with $\varepsilon_t = rL_t^c$ and $\delta_t = r(1+L_t^P)$. Therefore, the bound in Theorem 2 can be explicitly written as

$$W_t^{\mathcal{P},\mu^{\mathcal{E}}}(h_t) - W_t^{\mathcal{P}}(h_t) \le 2r\left[L_t^c + \sum_{\tau=t}^{T-1}\big[(1+L_t^P)\operatorname{Lip}(V_{t+1}^{\mathcal{M}}) + L_{\tau+1}^c\big]\right]$$

where $\mu_t^{\mathcal{E}}(h_t) = \pi_t^{\mathcal{M}}(y_t)$. This bound scales linearly with r, which means that as the observation becomes closer to the underlying state, the performance of the generalized certainty equivalent policy approaches that of the optimal policy.

B. Example 2

For the model presented in Example 2,

$$\begin{aligned}&\mathbb{E}[d_{\mathcal{S}}(S_t, \mathcal{E}_t(H_t))|h_t] \\ &\quad= (1-p)\mathbb{E}\big[d_{\mathcal{S}}(S_t,S_t)|h_t\big] + p\mathbb{E}\big[d_{\mathcal{S}}(S_t,U_t)|h_t\big] \\ &\quad\le p \sup_{s' \in \mathcal{S}} \mathbb{E}\big[d_{\mathcal{S}}(S_t,s')|h_t\big] \le pD\end{aligned}$$

where $D := \sup_{s_1,s_2 \in \mathcal{S}} d_{\mathcal{S}}(s_1,s_2)$ is the diameter of the space $\mathcal{S}$. Hence, we have $\eta_t \le pD$. Thus, if the model $\mathcal{M}$ satisfies Assumptions 1 and 2, then $(\mathcal{E}, c, P)$ is an AIS generator with $\varepsilon_t = pDL_t^c$ and $\delta_t = pD(1+L_t^P)$. Therefore, the bound in Theorem 2 can be explicitly written as

$$W_t^{\mathcal{P},\mu^{\mathcal{E}}}(h_t) - W_t^{\mathcal{P}}(h_t) \le 2pD\left[L_t^c + \sum_{\tau=t}^{T-1}\big[(1+L_t^P)\operatorname{Lip}(V_{t+1}^{\mathcal{M}}) + L_{\tau+1}^c\big]\right].$$

where $\mu_t^{\mathcal{E}}(h_t) = \pi_t^{\mathcal{M}}(y_t)$. This bound scales linearly with p, the probability of receiving a random observation. As p approaches zero, i.e., as the probability of correctly observing the state increases, the bound decreases, indicating that the

generalized certainty equivalent policy becomes closer to the optimal policy.

These results demonstrate that when the state estimation error is small, either due to observations with small noise (Example 1) or observations that are frequently accurate (Example 2), generalized 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. Example 3

For the MDP learning setting of Example 3, suppose

$$\mathbb{E}[d_{\mathcal{S}}(S_t, \mathcal{E}_t(H_t))|h_t] = \mathbb{E}[d_\Theta(\theta_t, \hat{\theta}_t)|h_t] \le \eta_t. \tag{13}$$

Suppose there exist finite constants $\{L_t^c\}_{t=1}^T$ and $\{L_t^P\}_{t=1}^T$ such that for any $\tilde{s} \in \tilde{\mathcal{S}}$, $a \in \mathcal{A}$ and any $\theta_1, \theta_2 \in \Theta$, we have

$$\begin{aligned}\big|c_t(\tilde{s}_1, a; \theta_1) - c_t(\tilde{s}_2, a; \theta_2)\big| &\le L_t^c(d_{\tilde{\mathcal{S}}}(\tilde{s}_1, \tilde{s}_2) + d_\Theta(\theta_1, \theta_2)),\\ d_{\mathrm{Was}}(P_{\tilde{S},t}(\cdot|\tilde{s}_1, a; \theta_1), P_{\tilde{S},t}(\cdot|\tilde{s}_2, a; \theta_2)) &\\ \le L_t^P(d_{\tilde{\mathcal{S}}}(\tilde{s}_1, \tilde{s}_2) + d_\Theta(\theta_1, \theta_2)).&\end{aligned}$$

Then, Assumptions 2 is satisfied with $\{L_t^c\}_{t=1}^T$ and $\{L_t^P\}_{t=1}^T$. As a result, under Assumption 1, the bound in Theorem 2 also holds in the MDP learning setting with η_t being the parameter estimation error given by (13).

D. Example 4

In Example 4 assume that the per-step cost is Lipschitz and satisfies (6). It can be shown that the dynamics are Lipschitz and satisfy (7) with $L_t^P = 1$. Further,

$$\mathbb{E}\big[|S_t - \mathcal{E}(h_t)| \;\big|\; h_t\big] = \mathbb{E}\left[\left|\frac{1}{t}\sum_{\tau=1}^{t} N_\tau\right| \;\middle|\; h_t\right] \le r.$$

Hence, we have $\eta_t \le r$. Thus, if the model $\mathcal{M}$ satisfies Assumptions 1 and 2, then $(\mathcal{E}, c, P)$ is an AIS generator with $\varepsilon_t = rL_t^c$ and $\delta_t = r(1 + L_t^P) = 2r$. Therefore, the bound in Theorem 2 can be explicitly written as

$$W_t^{\mathcal{P},\mu^{\mathcal{E}}}(h_t) - W_t^{\mathcal{P}}(h_t) \le 2r\left[L_t^c + \sum_{\tau=t}^{T-1}\big[2\,\mathrm{Lip}(V_{t+1}^{\mathcal{M}}) + L_{\tau+1}^c\big]\right].$$

VII. Conclusion

In this paper, we introduced a generalization of the certainty equivalence principle for control policies in partially observable Markov decision processes (POMPDs). 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 [12] to quantify the impact of estimation errors on control performance, deriving bounds in terms of the Lipschitz constants of the system dynamics and the per-step cost function.

To illustrate the practical relevance of our results, we examined several examples, including settings where the observation noise is small, cases with partial state observations, and learning scenarios where the system model is learned. These examples demonstrated that generalized certainty equivalent policies can perform near-optimally when state estimation errors are small. The bounds we derived provide quantitative measures of sub-optimality that scale linearly with the estimation error, highlighting that as observations become more accurate, the performance of generalized certainty equivalent policies approaches that of optimal policies. This suggests that in scenarios where exact optimal policies are computationally intractable, generalized 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] H. Theil, "Econometric models and welfare maximization," *Wirtschaftliches Archiv*, vol. 72, pp. 60–83, 1954.

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

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] A. D. Kara, "Near optimality of finite memory feedback policies in partially observed markov decision processes," *Journal of Machine Learning Research*, vol. 23, no. 1, pp. 437–482, 2022.

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

[16] 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.

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