2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Worst-Case Control and Learning Using Partial Observations Over an Infinite Time Horizon

Aditya Dave[1], Ioannis Faros[2], Nishanth Venkatesh[2], *Student Members, IEEE,*
Andreas A. Malikopoulos[1,2], *Senior Member, IEEE*

***Abstract*— Safety-critical cyber-physical systems require control strategies whose worst-case performance is robust against adversarial disturbances and modeling uncertainties. In this paper, we present a framework for approximate control and learning in partially observed systems to minimize the worst-case discounted cost over an infinite time horizon. We model disturbances to the system as finite-valued uncertain variables with unknown probability distributions. For problems with known system dynamics, we construct a dynamic programming (DP) decomposition to compute the optimal control strategy. Our first contribution is to define information states that improve the computational tractability of this DP for a class of problems with observable incurred costs at each time instance. Our second contribution proposes approximate information states that can be constructed or learned directly from observed data for these problems. We derive bounds on the performance loss of the resulting approximate control strategy and illustrate the effectiveness of our approach in partially observed decision-making problems with a numerical example.**

I. Introduction

Cyber-physical systems, such as connected and automated vehicles [1], often require decision-making in uncertain environments with partial knowledge of the dynamics [2] over long time horizons. This decision-making challenge is typically modeled with a *stochastic formulation*, where an agent accesses a prior probability distribution for all uncertainties and computes a control strategy to minimize the expected value of a discounted total cost across an infinite time horizon [3]. However, the actual performance of such a strategy degrades when the assumed prior distribution is different from the actual underlying distribution [4]. To mitigate this drawback, research efforts have proposed alternatives, including *(1) robust stochastic formulations,* where an agent minimizes the worst-case expected cost given a set of feasible probability distributions [5]; and *(2) risk-averse formulations,* where an agent minimizes a combination of both the expected cost and the cost variance [6], [7]. While these formulations improve the performance under a distribution mismatch, many safety-critical applications require further guarantees on the worst-case performance of a strategy against either adversarial attacks or system failure, e.g., cyber-security [8] and cyber-physical networks [9]. A *non-stochastic formulation* is suitable in such applications, where the agent has no knowledge of the distributions on the uncertainties and uses only their set of feasible values to compute a control strategy that minimizes the *maximum* possible cost [10]–[15]. This non-stochastic formulation is both maximally robust and risk-averse [16].

In this paper, we analyze a non-stochastic problem over an infinite time horizon with an agent that can access only output data and may not know the underlying state-space model. In such partially observed problems, an optimal strategy can be computed using a memory-based dynamic program (DP) when the time horizon is finite. However, the agent's memory grows to an infinite-dimensional vector as time tends to infinity. This makes a memory-based approach computationally intractable. When the state-space model is known, this challenge is alleviated using the maximum cost-to-come as an *information state* in the DP, for both finite-time [17] and infinite-time problems [18]. The computational tractability of this DP has been further improved in finite-time problems using *approximate information states* [19], [20]. Meanwhile, general notions of information states and their approximations have been developed for stochastic problems over an infinite time horizon without relying on state-space models [21]. In contrast, to the best of our knowledge, no such general notions exist for infinite-time problems to minimize the worst-case discounted total cost.

Our main contributions in this paper are: (1) we introduce general information states (Definition 1) and a time-invariant DP to compute an optimal strategy in non-stochastic problems over an infinite time horizon for problems with observable costs (Theorem 1); (2) we define approximate information states (Definition 2) to compute a strategy with a bounded performance loss (Theorem 2); and (3) using a numerical example, we show that approximate information states can be learned directly from output data with incomplete access to system dynamics. Subsequently, we compute an approximate strategy using deep Q-learning (Section V).

The remainder of the paper proceeds as follows. In Section II, we present our formulation. In Section III, we define the memory-based DP. In Section IV, we derive information states for observable costs, define approximate information states, and derive performance bounds. In Section V, we demonstrate our results in a numerical example, and in Section VI, we draw concluding remarks.

A. Notation and Preliminaries

In our exposition, we use the mathematical framework of uncertain variables [22]. For a set $\mathscr{X}$, an uncertain variable is

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

[1] School of Civil and Environmental Engineering, Cornell University, Ithaca, NY 14850, USA.

[2] Systems Engineering, Cornell University, Ithaca, NY 14850, USA. Email: {a.dave, if74, ns942, amaliko}@cornell.edu

979-8-3503-0123-6/23/$31.00 ©2023 IEEE

a mapping $\mathsf{X} : \Omega \to \mathscr{X}$ and is compactly denoted by $\mathsf{X} \in \mathscr{X}$. This is the non-stochastic equivalent of a random variable. For any $\omega \in \Omega$, its realization is $\mathsf{X}(\omega) = \mathsf{x} \in \mathscr{X}$. Its *marginal range* is the set of feasible realizations $[[\mathsf{X}]] := \{\mathsf{X}(\omega) \,|\, \omega \in \Omega\} \subseteq \mathscr{X}$. The *joint range* of two uncertain variables $\mathsf{X} \in \mathscr{X}$ and $\mathsf{Y} \in \mathscr{Y}$ is the set of feasible simultaneous realizations $[[\mathsf{X}, \mathsf{Y}]] := \big\{\big(\mathsf{X}(\omega), \mathsf{Y}(\omega)\big) \,|\, \omega \in \Omega\big\} \subseteq \mathscr{X} \times \mathscr{Y}$. The two uncertain variables are *independent* if $[[\mathsf{X}, \mathsf{Y}]] = [[\mathsf{X}]] \times [[\mathsf{Y}]]$. The *conditional range* of X given a realization y of Y is the set $[[\mathsf{X}|\mathsf{y}]] := \big\{\mathsf{X}(\omega) \mid \mathsf{Y}(\omega) = \mathsf{y},\ \omega \in \Omega\big\}$.

Consider two bounded, non-empty subsets $\mathscr{X}, \mathscr{Y}$ of a metric space $(\mathscr{S}, \eta)$ with metric $\eta : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}_{\geq 0}$. The Hausdorff distance is the pseudo-metric

$$\mathcal{H}(\mathscr{X}, \mathscr{Y}) := \max\big\{\sup_{\mathsf{x} \in \mathscr{X}} \inf_{\mathsf{y} \in \mathscr{Y}} \eta(\mathsf{x}, \mathsf{y}), \sup_{\mathsf{y} \in \mathscr{Y}} \inf_{\mathsf{x} \in \mathscr{X}} \eta(\mathsf{x}, \mathsf{y})\big\}. \quad (1)$$

Furthermore, if $f : \mathscr{S} \to \mathbb{R}$ is a Lipschitz continuous function with a constant $L_f \in \mathbb{R}_{\geq 0}$, then [15, Lemma 5]:

$$\Big|\sup_{x \in \mathscr{X}} f(x) - \sup_{y \in \mathscr{Y}} f(y)\Big| \leq L_f \cdot \mathcal{H}(\mathscr{X}, \mathscr{Y}). \quad (2)$$

II. Problem Formulation

We consider the control of an uncertain system that evolves in discrete time steps. At each time $t \in \mathbb{N} = \{0, 1, 2, \dots\}$ an agent collects an observation of the system as the uncertain variable $Y_t \in \mathcal{Y}$ and generates a control action denoted by the uncertain variable $U_t \in \mathcal{U}$. After generating the action at each t, the agent incurs a cost denoted by the uncertain variable $C_t \in \mathcal{C} \subset \mathbb{R}_{\geq 0}$. The set $\mathcal{C}$ is bounded by $\min\{\mathcal{C}\} = c^{\min}$ and $\max\{\mathcal{C}\} = c^{\max}$. We formulate our problem for a general case where the agent may not have knowledge of a state-space model for the system. Thus, we use an *input-output model* to describe the evolution of the system, as follows. At each $t \in \mathbb{N}$, the system receives two inputs: the action U_t, and an uncontrolled disturbance $W_t \in \mathcal{W}$. The disturbances $\{W_t \,|\, t \in \mathbb{N}\}$ constitute a sequence of independent uncertain variables. After receiving the inputs at each time $t \in \mathbb{N}$, the system generates two outputs: (1) the observation $Y_{t+1} = h_{t+1}(W_{0:t}, U_{0:t})$, where $h_{t+1} : \mathcal{W}^t \times \mathcal{U}^t \to \mathcal{Y}$ is the observation function; and (2) the cost $C_t = d_t(W_{0:t}, U_{0:t})$, where $d_t : \mathcal{W}^t \times \mathcal{U}^t \to \mathcal{C}$ is the cost function. The initial observation is $Y_0 = h_0(W_0)$.

The agent perfectly recalls all observations and control actions and at each $t \in \mathbb{N}$, the agent's memory is the uncertain variable $M_t := (Y_{0:t}, U_{0:t-1})$ taking values in $\mathcal{M}_t := \mathcal{Y}^t \times \mathcal{U}^{t-1}$. The agent uses a control law $g_t : \mathcal{M}_t \to \mathcal{U}$ to generate the action $U_t = g_t(M_t)$ as a function of the memory. The control strategy is the collection of control laws $\boldsymbol{g} := (g_0, g_1, \dots)$ with a feasible set $\mathcal{G}$. The performance of a strategy $\boldsymbol{g} \in \mathcal{G}$ is given by the *worst-case discounted cost*,

$$\mathcal{J}(\boldsymbol{g}) := \lim_{T \to \infty} \sup_{c_{0:T} \in [[C_{0:T}]]^{\boldsymbol{g}}} \sum_{t=0}^{T} \gamma^t \cdot c_t, \quad (3)$$

where $\gamma \in (0, 1)$ is a discount parameter, the marginal range $[[C_{0:T}]]^{\boldsymbol{g}}$ is the set of all feasible costs consistent with the strategy $\boldsymbol{g}$ and with the set of feasible disturbances $\mathcal{W}$. The limit in (3) is well defined because $C_t \leq c^{\max}$ for all t. Next, we define the control problem with known dynamics.

Problem 1. The optimization problem is to derive the infimum value $\inf_{\boldsymbol{g} \in \mathcal{G}} \mathcal{J}(\boldsymbol{g})$, given the feasible sets $\{\mathcal{U}, \mathcal{W}, \mathcal{Y}, \mathcal{C}\}$ and the functions $\{h_t, d_t \,|\, t \in \mathbb{N}\}$.

If the minimum value is achieved in Problem 1, the argument $\boldsymbol{g}^* = \arg\min_{\boldsymbol{g} \in \mathcal{G}} \mathcal{J}(\boldsymbol{g})$ is called an optimal control strategy. We aim to tractably compute an optimal strategy. We impose the following assumption in our analysis.

Assumption 1. We consider that the sets $\{\mathcal{U}, \mathcal{W}, \mathcal{Y}\}$ are each bounded subsets of a metric space $(\mathscr{S}, \eta)$ and $\mathcal{C}$ is a bounded subset of $\mathbb{R}_{\geq 0}$.

Assumption 1 ensures that all uncertain variables take values in bounded sets and that we can use the Hausdorff pseudo-metric (1) as a distance measure between them. In the next section, we derive a DP for Problem 1 with known dynamics. However, our main results in Section IV are also suitable for reinforcement learning problems with unknown dynamics. We illustrate this with an example in Section V.

III. Memory-based Dynamic Program

In this section, we present value functions to evaluate the performance of any strategy $\boldsymbol{g} \in \mathcal{G}$. Next, we present a memory-based DP decomposition of Problem 1 that approximately computes the value functions with arbitrary precision. Then, we highlight the drawback of the memory-based DP, which motivates the notion of information states presented in Section IV. To construct value functions, we first define the *accrued cost* at each $t \in \mathbb{N}$ as the sum of past incurred costs $A_t := \sum_{\ell=0}^{t-1} \gamma^\ell \cdot C_\ell$, which satisfies $A_{t+1} = A_t + \gamma^t \cdot C_t$ with $A_0 := 0$. This is well defined in the limit $t \to \infty$ because $\lim_{t \to \infty} A_t \leq \lim_{t \to \infty} \sum_{\ell=0}^{t-1} \gamma^\ell \cdot c^{\max} = \frac{c^{\max}}{1-\gamma} =: a^{\max}$. Thus, $A_t \in [0, a^{\max}]$ for all $t \in \mathbb{N}$. Similarly, the *cost-to-go* at any $t \in \mathbb{N}$ is the sum of future all costs still to be incurred $C_t^\infty := \sum_{\ell=t}^{\infty} \gamma^{\ell-t} \cdot C_\ell$. Note that $C_t^\infty \in [0, a^{\max}]$ for all t and that $C_t^\infty = C_t + \gamma \cdot C_{t+1}^\infty$. Then, for all $t \in \mathbb{N}$, we can define a value function for any $\boldsymbol{g} \in \mathcal{G}$ as

$$V_t^{\boldsymbol{g}}(m_t) := \sup_{a_t, c_t^\infty \in [[A_t, C_t^\infty | m_t]]^{\boldsymbol{g}}} \left(a_t + \gamma^t \cdot c_t^\infty\right), \quad (4)$$

where $[[A_t, C_t^\infty | m_t]]^{\boldsymbol{g}}$ is the conditional range induced by the choice of strategy $\boldsymbol{g}$. From the definition of the value functions, at $t = 0$ it holds that $\sup_{y_0 \in \mathcal{Y}} V_0^{\boldsymbol{g}}(y_0) = \mathcal{J}(\boldsymbol{g})$, where $m_0 = y_0$. Thus, the value function $V_0^{\boldsymbol{g}}(y_0)$ evaluates the performance of any strategy $\boldsymbol{g}$ for an initial observation y_0. Similarly, the optimal value function at each $t \in \mathbb{N}$ is

$$V_t(m_t) := \inf_{\boldsymbol{g} \in \mathcal{G}} V_t^{\boldsymbol{g}}(m_t), \quad (5)$$

and the optimal value is $\inf_{\boldsymbol{g} \in \mathcal{G}} \mathcal{J}(\boldsymbol{g}) = \sup_{y_0 \in \mathcal{Y}} V_0(y_0)$.

Given the value functions in (4) and (5), we can evaluate the performance of a strategy and compare it with the optimal performance. However, there is no natural DP decomposition to compute these value functions in an infinite-horizon system with no terminal time. Thus, we construct a memory-based DP that assumes a finite horizon $T \in \mathbb{N}$ and use it to compute approximations of the value functions recursively. For any $\boldsymbol{g} \in \mathcal{G}$, we define finite-horizon evaluation functions for all $m_t \in \mathcal{M}_t$ and each $t = 0, \dots, T-1$ as $J_t^{\boldsymbol{g}}(m_t;T) := \sup_{m_{t+1} \in [[M_{t+1} | m_t]]^{\boldsymbol{g}}} J_{t+1}^{\boldsymbol{g}}(m_{t+1};T)$, where $J_T^{\boldsymbol{g}}(m_T;T) := \sup_{a_T, c_T \in [[A_T, C_T | m_T]]^{\boldsymbol{g}}} (a_T + \gamma^T \cdot c_T)$ at time T. Similarly, we define approximately optimal finite-horizon functions for all $m_t \in \mathcal{M}_t$ and each $t = 0, \dots, T-1$ as

$$J_t(m_t;T) := \inf_{u_t \in \mathcal{U}} \sup_{m_{t+1} \in [[M_{t+1} | m_t, u_t]]} J_{t+1}(m_{t+1};T), \quad (6)$$

where, $J_T(m_T;T) := \inf_{u_T \in \mathcal{U}} \sup_{a_T, c_T \in [[A_T, C_T | m_T, u_T]]} (a_T + \gamma^T \cdot c_T)$. Note that the finite-horizon functions $J_t^{\boldsymbol{g}}(m_t;T)$ and $J_t(m_t;T)$ at any $t = 0, \dots, T$ are parameterized by the choice of horizon $T \in \mathbb{N}$. Next, we bound the approximation error between the value functions and their finite-horizon counterparts.

Lemma 1. *For any finite horizon $T \in \mathbb{N}$ and for all $m_t \in \mathcal{M}_t$ and each $t = 0, \dots, T$,*

$$\textit{a)}\ \frac{\gamma^{T+1} \cdot c^{\min}}{1-\gamma} + J_t^{\boldsymbol{g}}(m_t;T) \leq V_t^{\boldsymbol{g}}(m_t) \leq J_t^{\boldsymbol{g}}(m_t;T) + \frac{\gamma^{T+1} \cdot c^{\max}}{1-\gamma}, \quad (7)$$

$$\textit{b)}\ \frac{\gamma^{T+1} \cdot c^{\min}}{1-\gamma} + J_t(m_t;T) \leq V_t(m_t) \leq J_t(m_t;T) + \frac{\gamma^{T+1} \cdot c^{\max}}{1-\gamma}. \quad (8)$$

Proof. a) We prove each inequality in (7) using backward induction. For the upper bound at time T, we use the dynamics of the accrued cost and cost-to-go to write that $V_T^{\boldsymbol{g}}(m_T) = \sup_{a_T, c_T, c_{T+1}^{\infty} \in [[A_T, C_T, C_{T+1}^{\infty} | m_T]]^{\boldsymbol{g}}} (a_T + \gamma^T \cdot c_T + \gamma^{T+1} \cdot c_{T+1}^{\infty}) \leq \sup_{a_T, c_T \in [[A_T, C_T | m_T]]^{\boldsymbol{g}}} (a_T + \gamma^T \cdot c_T) + \gamma^{T+1} \cdot a^{\max} \leq J_t^{\boldsymbol{g}}(m_t;T) + \frac{\gamma^{T+1} \cdot c^{\max}}{1-\gamma}$. The lower bound at time T follows from $\frac{c^{\min}}{1-\gamma} \leq c_{T+1}^{\infty}$ using the same sequence of arguments as before. This forms the basis of our induction. Next, consider the hypothesis that (7) holds at time $t+1$. For the upper bound at time t, by definition $V_t^{\boldsymbol{g}}(m_t) = \sup_{a_t, c_t, c_{t+1}^{\infty} \in [[A_t, C_t, C_{t+1}^{\infty} | m_t]]^{\boldsymbol{g}}} (a_t + \gamma^t \cdot c_t + \gamma^{t+1} \cdot c_{t+1}^{\infty}) = \sup_{a_{t+1}, c_{t+1}^{\infty} \in [[A_{t+1}, C_{t+1}^{\infty} | m_t]]^{\boldsymbol{g}}} (a_{t+1} + \gamma^{t+1} \cdot c_{t+1}^{\infty}) = \sup_{m_{t+1} \in [[M_{t+1} | m_t]]^{\boldsymbol{g}}} \sup_{a_{t+1}, c_{t+1}^{\infty} \in [[A_{t+1}, C_{t+1}^{\infty} | m_{t+1}]]^{\boldsymbol{g}}} (a_{t+1} + \gamma^{t+1} \cdot c_{t+1}^{\infty}) = \sup_{m_{t+1} \in [[M_{t+1} | m_t]]^{\boldsymbol{g}}} V_{t+1}^{\boldsymbol{g}}(m_{t+1}) \leq \sup_{m_{t+1} \in [[M_{t+1} | m_t]]^{\boldsymbol{g}}} J_{t+1}^{\boldsymbol{g}}(m_{t+1};T) + \frac{\gamma^{T+1} \cdot c^{\max}}{1-\gamma} = J_t^{\boldsymbol{g}}(m_t;T) + \frac{\gamma^{T+1} \cdot c^{\max}}{1-\gamma}$, where, in the fourth equality, we use (4) for $V_{t+1}^{\boldsymbol{g}}(m_{t+1})$; and in the inequality, we use the hypothesis. The lower bound follows from the same sequence of arguments. Thus, (7) holds using induction.

b) We can prove the lower bound in (8) by taking the infimum on both sides of the lower bound in (7). To prove the upper bound in (8), we first note that $J_t(m_t;T) = \inf_{\boldsymbol{g} \in \mathcal{G}} J_t^{\boldsymbol{g}}(m_t;T)$ for all $t = 0, \dots, T$ using standard DP arguments for terminal-cost problems [17]. Then, at time T, by definition $V_T(m_T) = \inf_{\boldsymbol{g} \in \mathcal{G}} V_T^{\boldsymbol{g}}(m_T) \leq \inf_{\boldsymbol{g} \in \mathcal{G}} J_T^{\boldsymbol{g}}(m_T;T) + \frac{\gamma^{T+1} \cdot c^{\max}}{1-\gamma} = J_T(m_T;T) + \frac{\gamma^{T+1} \cdot c^{\max}}{1-\gamma}$. Using this as the basis, the result follows for all $t = 0, \dots, T$ using the same induction arguments as in (7). □

Lemma 1 establishes that the approximation error between finite-horizon functions and corresponding value functions decreases with the horizon $T \in \mathbb{N}$. As a direct consequence of (8), $\lim_{T \to \infty} J_0(y_0;T) = V_0(y_0)$ for all $y_0 \in \mathcal{Y}$. However, the domain of $J_T(m_T;T)$ is $\mathcal{M}_T = \mathcal{Y}^T \times \mathcal{U}^{T-1}$ which grows with T, and in the limit $T \to \infty$, the set $\mathcal{M}_T$ is infinite-dimensional. Thus, it is computationally intractable to achieve close approximations of the optimal value using (6). We address this issue in the next section using *information states*, which take values in time-invariant spaces.

IV. Systems with Observable Costs

In this section, we analyze Problem 1 in the case where the agent observes the incurred cost at each instance of time. Thus, at each $t \in \mathbb{N}$, the agent receives a realization of (Y_t, C_t) and the memory is $M_t = (Y_{0:t}, C_{0:t-1}, U_{0:t-1})$. For such systems, we present the notion of information states which take values in time-invariant spaces. Then, we use them to construct a time-invariant DP decomposition, which converges to the optimal value of Problem 1. To begin, we formulate our notion of information states.

Definition 1. An *information state* for Problem 1 with observable costs at any $t \in \mathbb{N}$ is an uncertain variable $\bar{S}_t = \bar{\sigma}_t(M_t)$ taking values in a bounded, time-invariant set $\bar{\mathcal{S}}$. For all $t \in \mathbb{N}$, for all $m_t \in \mathcal{M}_t$ and $u_t \in \mathcal{U}_t$, it satisfies that

$$[[C_t, \bar{S}_{t+1} \,|\, m_t, u_t]] = [[C_t, \bar{S}_{t+1} \,|\, \bar{\sigma}_t(m_t), u_t]]. \quad (9)$$

Next, we use the information state from Definition 1 to construct a time-invariant operator $\bar{\mathcal{T}} : [\bar{\mathcal{S}} \to \mathbb{R}] \to [\bar{\mathcal{S}} \to \mathbb{R}]$ that yields a fixed-point equation to recursively compute the optimal value in Problem 1. For any uniformly bounded function $\bar{\Lambda} : \bar{\mathcal{S}} \to \mathbb{R}$, the operator $\bar{\mathcal{T}}$ is given by

$$[\bar{\mathcal{T}} \bar{\Lambda}](\bar{s}) := \inf_{u \in \mathcal{U}} \sup_{c, \bar{s}' \in [[C, \bar{S}' | \bar{s}, u]]} \big(c + \gamma \cdot \bar{\Lambda}(\bar{s}')\big), \quad (10)$$

for all $s \in \mathcal{S}$. Note that we use time-invariant notation for all variables in (10) because the sets and functions in the RHS are time-invariant. Due to discounting, $\mathcal{T}$ is a contraction mapping. Thus, using the Banach fixed point theorem, the equation $\bar{\Lambda} = \bar{\mathcal{T}} \bar{\Lambda}$ admits a unique solution $\bar{\Lambda}^{\infty} = \bar{\mathcal{T}} \bar{\Lambda}^{\infty}$. Starting with $\bar{\Lambda}^0(\bar{s}) := 0$, the fixed-point iteration around $\bar{\mathcal{T}}$ generates a sequence of functions

$$\bar{\Lambda}^{n+1}(\bar{s}) = [\bar{\mathcal{T}} \bar{\Lambda}^n](\bar{s}) = [\bar{\mathcal{T}}^n \bar{\Lambda}^0](\bar{s}), \quad (11)$$

for all $n = 1, 2, \dots$, such that $\lim_{n \to \infty} \bar{\mathcal{T}}^n \bar{\Lambda}^0 = \bar{\Lambda}^{\infty}$. Next, we that $\bar{\Lambda}^n(\bar{\sigma}_t(m_t))$, $n \in \mathbb{N}$ can be used to estimate $V_t(m_t)$ for all t in Problem 1 with observable costs, with an error that decreases in n.

Theorem 1. *Consider the function $\bar{\Lambda}^n$ generated using* (11) *for any $n \in \mathbb{N}$. Then, for all $t \in \mathbb{N}$, it holds that*

$$\frac{\gamma^{n+t} \cdot c^{\min}}{1-\gamma} + \gamma^t \cdot \bar{\Lambda}^n(\bar{\sigma}_t(m_t)) + \sup_{a_t \in [[A_t | m_t]]} a_t \ \leq \ V_t(m_t)$$

$$\leq \sup_{a_t \in [[A_t|m_t]]} a_t + \gamma^t \cdot \bar{\Lambda}^n(\bar{\sigma}_t(m_t)) + \frac{\gamma^{n+t} \cdot c^{\max}}{1-\gamma}. \quad (12)$$

Proof. We show (12) by combining arguments in Lemma 1 with the definition of information states in (9). Thus, we first show that for any horizon $T \in \mathbb{N}$, the following relationship is true for each $t = 0, \dots, T$:

$$J_t(m_t; T) = \gamma^t \cdot \bar{\Lambda}^{T-t+1}\big(\bar{\sigma}_t(m_t)\big) + \sup_{a_t \in [[A_t|m_t]]} a_t. \quad (13)$$

We can prove (13) by induction. At time T, using the definition of the finite-horizon function $J_T(m_T; T) = \inf_{u_T \in \mathcal{U}} \sup_{a_T, c_T \in [[A_T, C_T|m_T, u_T]]}(a_T + \gamma^T \cdot c_T) = \inf_{u_T \in \mathcal{U}} \sup_{c_T \in [[C_T|m_T, u_T]]} c_T + \sup_{a_T \in [[A_T|m_T]]} a_T = \inf_{u_T \in \mathcal{U}} \sup_{c_T, \bar{\sigma}_{T+1}(m_{T+1}) \in [[C_T, S_{T+1}|m_T, u_T]]}(c_T + \gamma^T \cdot \bar{\Lambda}^0(\bar{\sigma}_{T+1}(m_{T+1}))) + \sup_{a_T \in [[A_T|m_T]]} a_T = \bar{\Lambda}^1(\sigma_T(m_T)) + \sup_{a_T \in [[A_T|m_T]]} a_T$, where, in the second equality, note that A_T is completely determined given M_T because the costs are perfectly observed; and in the third equality, we note that $\bar{\Lambda}^0(\bar{\sigma}_{T+1}(m_{T+1})) = 0$. This forms the basis of our induction. Next, consider the hypothesis that (13) holds at time $t+1$. Using the definition of the finite-horizon function at time t, $J_t(m_t; T) = \inf_{u_t \in \mathcal{U}} \sup_{m_{t+1} \in [[M_{t+1}|m_t, u_t]]} \sup_{c_t, a_t \in [[C_t, A_t|m_{t+1}]]}(\gamma^{t+1} \cdot \bar{\Lambda}^{T-t}(\bar{\sigma}_{t+1}(m_{t+1})) + \gamma^t \cdot c_t + a_t) = \inf_{u_t \in \mathcal{U}} \sup_{c_t, a_t, m_{t+1} \in [[C_t, A_t, M_{t+1}|m_t, u_t]]}(\gamma^{t+1} \cdot \bar{\Lambda}^{T-t}(\bar{\sigma}_{t+1}(m_{t+1})) + \gamma^t \cdot c_t + a_t) = \inf_{u_t \in \mathcal{U}} \sup_{c_t, \bar{\sigma}_{t+1}(m_{t+1}) \in [[C_t, \bar{S}_{t+1}|m_t, u_t]]} (\gamma^{t+1} \cdot \bar{\Lambda}^{T-t}(\bar{\sigma}_{t+1}(m_{t+1})) + \gamma^t \cdot c_t) + \sup_{a_t \in [[A_t|m_t]]} a_t = \inf_{u_t \in \mathcal{U}} \sup_{c_t, \bar{\sigma}_{t+1}(m_{t+1}) \in [[C_t, \bar{S}_{t+1}|\bar{\sigma}_t(m_t), u_t]]}(\gamma^{t+1} \cdot \bar{\Lambda}^{T-t}(\bar{\sigma}_{t+1}(m_{t+1})) + \gamma^t \cdot c_t) + \sup_{a_t \in [[A_t|m_t]]} a_t = \gamma^t \cdot \bar{\Lambda}^{T-t+1}(\bar{\sigma}_t(m_t)) + \sup_{a_t \in [[A_t|m_t]]} a_t$, where, in the third equality, we use the fact that costs are observable; in the fourth equality, we use (9) from Definition 1; and in the last equality, we use the definition of $\bar{\Lambda}^{T-t+1}$ from (11). This proves (13) using induction. Then, (12) follows directly for all $t \in \mathbb{N}$ and all $n \in \mathbb{N}$ by substituting (13) into (8) and selecting a horizon $T = t + n - 1$. □

Theorem 1 allows us to characterize the error between the optimal value $V_0(y_0)$ and the value $\bar{\Lambda}^n(\bar{\sigma}_0(y_0))$ for any $y_0 \in \mathcal{Y}$, by selecting $t = 0$ in (12). Then, we select $t = 0$ and let $n \to \infty$ to establish that $\bar{\Lambda}^\infty(\bar{\sigma}_0(y_0)) = V_0(y_0)$. Thus, when Problem 1 has observable costs, the fixed point $\bar{\Lambda}^\infty$ computes the optimal value function V_0 as a direct consequence of Theorem 1. Next, consider that the infimum is achieved in the RHS of $[\bar{\mathcal{T}}\bar{\Lambda}^n](\bar{s})$ for all $\bar{s} \in \bar{\mathcal{S}}$ and $n \in \mathbb{N}$. We define a strategy $\boldsymbol{\pi}^* = (\pi^*, \pi^*, \dots)$, where $\bar{\pi}^* : \bar{\mathcal{S}} \to \mathcal{U}$ is the minimizing argument in RHS of (10) for $\Lambda = \bar{\Lambda}^\infty$. Then, it holds that the memory-based strategy $\bar{\boldsymbol{g}}^* = (\bar{g}_0^*, \bar{g}_1^*, \dots)$, where $\bar{g}_t^* := \bar{\pi}^*(\sigma_t(m_t))$, gives an optimal solution to Problem 1 with observable costs (from standard arguments analogous to Appendix B of [23]).

Remark 1. For systems with perfectly observed states and known dynamics, an example of a valid information state at any time t is the state itself $X_t \in \mathcal{X}$. For partially observed systems, a valid information state at any time t is the conditional range $[[X_t|M_t]] \in 2^{\mathcal{X}}$.

Remark 2. When attempting to learn an information state that satisfies Definition 1 using only output data, we may not be able to satisfy (9) exactly. Thus, in Subsection IV-A, we relax this definition for approximate information states.

A. Approximate Information States

In this subsection, we define approximate information states that approximately satisfy (9), and construct a time-invariant approximate DP of Problem 1 using them. Then, we bound the resulting error, estimating the optimal value and the performance loss of the resulting approximate strategy.

Definition 2. An *approximate information state* for Problem 1 with observable costs at any $t \in \mathbb{N}$ is an uncertain variable $\hat{S}_t = \hat{\sigma}_t(M_t)$ taking values in a bounded, time-invariant set $\hat{\mathcal{S}}$. Furthermore, there exists a parameter $\epsilon \in \mathbb{R}_{\geq 0}$ such that for all $m_t \in \mathcal{M}_t$ and $u_t \in \mathcal{U}$ and $t \in \mathbb{N}$, it satisfies

$$\mathcal{H}\big([[C_t, \hat{S}_{t+1}|m_t, u_t]], [[C_t, \hat{S}_{t+1}|\hat{\sigma}_t(m_t), u_t]]\big) \leq \epsilon, \quad (14)$$

where recall that $\mathcal{H}$ is the Hausdorff distance defined in (1).

To compute an approximate value and control strategy, we proceed with approximate information states just as we did with information states. First, we construct a time-invariant operator $\hat{\mathcal{T}} : [\hat{\mathcal{S}} \to \mathbb{R}] \to [\hat{\mathcal{S}} \to \mathbb{R}]$, such that for any uniformly bounded function $\hat{\Lambda} : \hat{\mathcal{S}} \to \mathbb{R}$,

$$[\hat{\mathcal{T}}\hat{\Lambda}](\hat{s}) := \inf_{u \in \mathcal{U}} \sup_{c, \hat{s}' \in [[C, \hat{S}'|\hat{s}, u]]} \big(c + \gamma \cdot \hat{\Lambda}(\hat{s}')\big). \quad (15)$$

Note that $\hat{\mathcal{T}}$ is a contraction mapping and thus, the equation $\hat{\Lambda} = \hat{\mathcal{T}}\hat{\Lambda}$ admits a unique solution $\hat{\Lambda}^\infty = \hat{\mathcal{T}}\hat{\Lambda}^\infty$. Then, starting with $\hat{\Lambda}^0(\hat{s}) := 0$ the fixed-point iteration around $\hat{\mathcal{T}}$ recursively generates the functions $\hat{\Lambda}^{n+1}(\hat{s}) = [\hat{\mathcal{T}}\hat{\Lambda}^n](\hat{s}) = [\hat{\mathcal{T}}^n\hat{\Lambda}^0](\hat{s})$, for all $n = 1, 2, \dots$, such that $\lim_{n \to \infty} \hat{\mathcal{T}}^n \hat{\Lambda}^0 = \hat{\Lambda}^\infty$. This forms our approximate DP decomposition. Next, consider that the infimum is achieved in the RHS of $[\hat{\mathcal{T}}\hat{\Lambda}^n](\hat{s})$ for all $\hat{s} \in \hat{\mathcal{S}}$ and all $n \in \mathbb{N}$. We an approximate strategy $\hat{\boldsymbol{\pi}}^* = (\hat{\pi}^*, \hat{\pi}^*, \dots)$, where $\hat{\pi}^* : \hat{S} \to \mathcal{U}$ is the minimizing argument in the RHS of (15) for $\hat{\Lambda} = \hat{\Lambda}^\infty$. Then, a corresponding memory-based strategy is $\hat{\boldsymbol{g}}^* := (\hat{g}_0^*, \hat{g}_1^*, \dots)$ with $\hat{g}_t^* := \pi^*(\sigma_t(m_t))$ for all $t \in \mathbb{N}$. Next, we bound both the approximation error between the optimal value $V_0(y_0)$ and $\hat{\Lambda}^\infty(\hat{\sigma}_0(y_0))$, and the performance loss when implementing $\hat{\boldsymbol{g}}^*$ to generate the control actions.

Theorem 2. *Let the functions $\hat{\Lambda}^n$ be Lipschitz continuous with a constant $L_{\hat{\Lambda}} \in \mathbb{R}_{\geq 0}$ for all $n \in \mathbb{N}$. Then, we have that*

a) $|V_0(y_0) - \hat{\Lambda}^\infty(\hat{\sigma}_0(y_0))| \leq \hat{L} \cdot \epsilon \cdot (1-\gamma)^{-1}$, (16)

b) $|V_0(y_0) - V_0^{\hat{\boldsymbol{g}}^*}(y_0)| \leq 2 \cdot \hat{L} \cdot \epsilon \cdot (1-\gamma)^{-1}$, (17)

where $\hat{L} = \max\{\gamma \cdot L_{\hat{\Lambda}}, 1\}$.

Proof. We show (16) using (8) from Lemma 1. Thus, we first show that for any $T \in \mathbb{N}$, it holds for all $t = 0, \dots, T$:

$$|J_t(m_t; T) - \gamma^t \cdot \hat{\Lambda}^{T-t+1}\big(\hat{\sigma}_t(m_t)\big) - \sup_{a_t \in [[A_t|m_t]]} a_t| \leq \beta_t(T), \quad (18)$$

where $\beta_t(T) = \beta_{t+1}(T) + \gamma^t \cdot \hat{L} \cdot \epsilon$ and $\beta_T(T) = \gamma^T \cdot \hat{L} \cdot \epsilon$. We prove (18) by induction. At time T, recall that $J_T(m_T; T) =$

$\inf_{u_T\in\mathcal{U}}\sup_{c_T\in[[C_T|m_T,u_T]]}\gamma^T\cdot c_T + \sup_{a_T\in[[A_T|m_T]]} a_T$ for perfectly observed costs. This implies that $|J_T(m_T;T) - \gamma^T\cdot\hat{\Lambda}^1\big(\hat{\sigma}_T(m_T)\big) - \sup_{a_T\in[[A_T|m_T]]} a_T| = |\inf_{u_T\in\mathcal{U}}\sup_{c_T\in[[C_T|m_T,u_T]]}\gamma^T\cdot c_T - \gamma^T\cdot\hat{\Lambda}^1\big(\hat{\sigma}_T(m_T)\big)| = \gamma^T\cdot|\inf_{u_T\in\mathcal{U}}\sup_{c_T,\hat{s}_{T+1}\in[[C_T,\hat{S}_{T+1}|m_T,u_T]]}(c_T + \gamma\cdot\hat{\Lambda}^0(\hat{s}_{T+1})) - \inf_{u_T\in\mathcal{U}}\sup_{c_T,\hat{s}_{T+1}\in[[C_T,\hat{S}_{T+1}|\hat{\sigma}_T(m_T),u_T]]}(c_T+\gamma\cdot\hat{\Lambda}^0(\hat{s}_{T+1}))| \leq \gamma^T\cdot\hat{L}\cdot\sup_{u_T\in\mathcal{U}}\mathcal{H}([[C_T,\hat{S}_{T+1}|m_T,u_T]], [[C_T,\hat{S}_{T+1}|\hat{\sigma}_T(m_T),u_T]]) \leq \gamma^T\cdot\hat{L}\cdot\epsilon$, where, in the second equality, we note that $\Lambda^0(\hat{s}_{T+1}) = 0$ identically; in the first inequality, we note that $\hat{L} = \max\{\gamma\cdot L_{\hat{\Lambda}}, 1\}$ is the Lipschitz constant of $(c_T + \gamma\cdot\hat{\Lambda}^0(\hat{s}_{T+1}))$ with respect to $(c_T, \hat{s}_{T+1})$ and use (2); and in the second inequality, we use (14). This forms the basis of our induction. Next, we consider the hypothesis that (16) holds at time $t+1$. Using the hypothesis and rearranging terms, $J_{t+1}(m_{t+1};T) \leq \beta_{t+1}(T) + \gamma^{t+1}\cdot\hat{\Lambda}^{T-t}\ (\hat{\sigma}_{t+1}(m_{t+1})) + \sup_{a_{t+1}\in[[A_{t+1}|m_{t+1}]]} a_{t+1}$. Then, at time t, $|J_t(m_t;T) - \gamma^t\cdot\hat{\Lambda}^{T-t+1}\big(\hat{\sigma}_t(m_t)\big) - \sup_{a_t\in[[A_t|m_t]]}\ a_t| \leq \beta_{t+1}(T) + |\inf_{u_t\in\mathcal{U}}\sup_{m_{t+1}\in[[M_{t+1}|m_t,u_t]]}(\gamma^{t+1}\cdot\hat{\Lambda}^{T-t}\ (\hat{\sigma}_{t+1}(m_{t+1}))+ \sup_{a_t,c_t\in[[A_t,C_t|m_{t+1}]]}\ (a_t+\gamma^t\cdot c_t)) - \gamma^t\cdot\ \hat{\Lambda}^{T-t+1}(\hat{\sigma}_t(m_t)) - \sup_{a_t\in[[A_t|m_t]]} a_t| \leq \beta_{t+1}(T) + \gamma^t\cdot\sup_{u_t\in\mathcal{U}} |\sup_{c_t,\hat{\sigma}_{t+1}(m_t)\in[[C_t,\hat{S}_{t+1}|m_t,u_t]]}(c_t+\gamma\cdot\hat{\Lambda}^{T-t}(\hat{\sigma}_{t+1}(\ m_{t+1}))) - \sup_{c_t,\hat{s}_{t+1}\in[[C_t,\hat{S}_{t+1}|\hat{\sigma}_t(m_t),u_t]]}(c_t + \gamma\cdot\hat{\Lambda}^{T-t}(\hat{s}_{t+1}))| \leq \beta_{t+1}(T) + \gamma^t\cdot\hat{L}\cdot\epsilon$, where, in the second inequality, we use $[[A_t,C_t|m_{t+1}]] = [[A_t|m_{t+1}]] \times [[C_t|m_{t+1}]]$ when a_t is perfectly observed; and, in the third inequality, we use (2) and (14). This proves (18) for all t using induction.

Next, for the iterated function $\hat{\Lambda}^n$, we select a horizon $T = n-1$ and set $t = 0$ in (18), to write that $|J_0(y_0;T) - \hat{\Lambda}^n(\hat{\sigma}_0(y_0))| \leq \beta_0(T)$, where $\beta_0(T) = \sum_{\ell=0}^{n-1}\gamma^\ell\cdot\hat{L}\cdot\epsilon$. As $n \to \infty$ with $T = n-1$, note that $\lim_{T\to\infty} J_0(y_0;T) = V_0(y_0)$, $\lim_{n\to\infty}\hat{\Lambda}^n(\hat{\sigma}_0(y_0)) = \hat{\Lambda}^\infty(\hat{\sigma}_0(y_0))$, and $\lim_{T\to\infty}\beta_0(T) = \frac{\hat{L}\cdot\epsilon}{1-\gamma}$. The proof for (17) follows from a similar series of arguments. □

B. Alternate Characterization

When exploring whether an uncertain variable is a valid candidate to be considered for an approximate information state, it may be difficult to verify (14). Thus, we present two *stronger* conditions that are easier to verify. To establish that $\hat{S}_t = \hat{\sigma}_t(M_t)$, $t \in \mathbb{N}$, satisfies (14), the following two conditions should hold (see proof in Appendix C of [23]):

1) State-like evolution: There exists a Lipschitz continuous function $\psi : \hat{\mathcal{S}} \times \mathcal{U} \times \mathcal{Y} \to \mathcal{S}$, such that

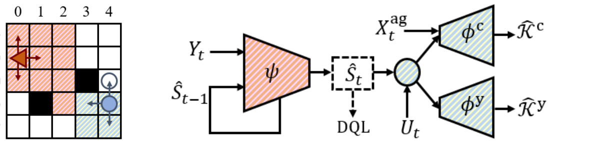
$$\hat{\sigma}_{t+1}(M_{t+1}) = \psi(\hat{\sigma}_t(M_t), U_t, Y_{t+1}). \tag{19}$$

2) Sufficient to approximate outputs: For all $m_t \in \mathcal{M}_t$ and $u_t \in \mathcal{U}$, there exists a constant $\delta \in \mathbb{R}_{\geq 0}$ such that

$$\mathcal{H}([[C_t, Y_{t+1}|m_t, u_t]]], [[C_t, Y_{t+1}|\hat{\sigma}_t(m_t), u_t]]) \leq \delta. \tag{20}$$

V. Numerical Example

We consider an agent pursuing a target across a 5×5 grid with obstacles. At each $t \in \mathbb{N}$, the agent's position X_t^{ag} and the target's position X_t^{ta} each take values in the set of grid cells $\mathcal{X} = \big\{(0,0),(0,1),\dots,(4,4)\big\} \setminus \mathcal{O}$, where $\mathcal{O} \subset \mathcal{X}$ is the set of obstacles. Let $\mathcal{W} = \{(-1,0),(1,0),(0,0),(0,1),(0,-1)\}$, $\mathcal{N} = \{(0,-1),(0,0),(0,1)\}$, and $\mathcal{U} = \mathcal{W} \times \{\xi\}$, where ξ denotes a "stop" action. Starting at $X_0^{\text{ta}} \in \mathcal{X}$, the target's position evolves as $X_{t+1}^{\text{ta}} = \delta(X_t^{\text{ta}} + W_t \in \mathcal{X})\cdot(X_t^{\text{ta}} + W_t) + (1 - \delta(X_t^{\text{ta}} + W_t \in \mathcal{X}))\cdot X_t^{\text{ta}}$, where $W_t \in \mathcal{W}$ and δ is returns 1 or 0 after checking the argument. At each t, the agent observes their own position perfectly and the target's position as $Y_t = \delta(X_t^{\text{ta}} + N_t \in \mathcal{X})\cdot(X_t^{\text{ta}} + N_t) + (1 - \delta(X_t^{\text{ta}} + N_t \in \mathcal{X}))\cdot X_t^{\text{ta}}$, where $N_t \in \mathcal{N}$. Then, the agent selects an action $U_t \in \mathcal{U}$, e.g., to move or stop. If the agent moves, i.e., $U_t \neq \xi$, then $X_{t+1}^{\text{ag}} = \delta(X_t^{\text{ag}} + U_t \in \mathcal{X})\cdot(X_t^{\text{ag}} + U_t) + (1 - \delta(X_t^{\text{ag}} + U_t \in \mathcal{X}))\cdot X_t^{\text{ag}}$. The agent incurs a cost $C_t = 2$. If the agent stops, i.e., $U_t = \xi$, they incur a terminal cost $10\cdot\eta(X_T^{\text{ta}}, X_T^{\text{ag}})$ for the L1 distance from the target. We illustrate this pursuit problem in Fig. 1(a), where the black cells are obstacles, the red triangle is the agent, the blue circle is the observation, and the blue disk is the target.



(a) The grid (b) Encoder-decoder architecture

Fig. 1. The pursuit problem with $x_0^{\text{ag}} = (0,1)$, $x_0^{\text{ta}} = (4,3)$ and $y_0 = (4,2)$ is in (a). The neural network architecture for the AIS is in (b).

We consider the pursuit problem when the agent is aware of their own dynamics but unaware of the observation model and target's dynamics. Thus, we train an approximate information state (AIS) model to learn a representation of the target's dynamics using observations, actions, and incurred costs to enforce (19) and (20). The AIS is generated by a neural network in an encoder-decoder architecture, as shown in Fig. 1(b). At each $t \in \mathbb{N}$, the *encoder* ψ receives as an input the observation Y_t and previous AIS $\hat{S}_{t-1}$ and generates $\hat{S}_t$. It consists of a linear layer of size $(2,4)$ with ReLU activation, followed by a gated recurrent unit (GRU) with a hidden state size of 4. The hidden state of the GRU constitutes the AIS $\hat{S}_t$ updated recurrently as $\hat{S}_t = \psi(\hat{S}_{t-1}, Y_t)$, thus enforcing (19). Note that our AIS is independent of the agent's position and action because the target moves independently from the agent. The decoder is comprised of two separate units, each of which is selected according to the action U_t. If $U_t = \xi$, we use the network ϕ^{c} which takes as an input the agent's position X_t^{ag} and the AIS $\hat{S}_t$ and generates a set of possible terminal costs $\hat{\mathcal{K}}^{\text{c}} := [[C_t|X_t^{\text{ag}}, \hat{S}_t]]$. This network comprises of two linear layers with dimensions $(6,16)$ and $(16,9)$, where the first layer has ReLU activation and the second has sigmoid activation. If $U_t \neq \xi$, we use the network ϕ^{y} which takes the AIS $\hat{S}_t$ as an input and generates the conditional range $\hat{\mathcal{K}}^{\text{y}} := [[Y_{t+1}|\hat{S}_t]]$. This network comprises of two linear layers with dimensions $(6,16)$ and $(16,23)$, where the first layer has ReLU activation and the second has sigmoid activation.

We train the entire model simultaneously using the outputs of the decoder. At each $t \in \mathbb{N}$, the training loss is given by the

Hausdorff distance between the one-hot encoded incoming data point, either C_t or Y_{t+1}, and the current predicted set. Since the Hausdorff distance is not differentiable, we adapt the distance-transform-based surrogate loss proposed in [24]. Note that we cannot observe the true underlying set and thus train the predictions against sampled data points to eventually learn the feasible sets. We train the network for 3×10^6 instances with a learning rate of 0.0003. In each instance, we randomly initialize the agent and target's positions from the pink and blue hatched cells in Fig. 1(a) and randomize all subsequent noises, disturbances, and actions.

Next, we utilize the trained encoder's output AIS and the agent's position as a state input to a deep Q-learning network (DQN) with two layers of $(6, 3)$ and $(3, 6)$ and a LeakyReLU activation each. We train this AIS-DQN using an exploratory policy for 3×10^6 instances with a learning rate of 0.0005 using a maximally risk-averse approach from [6] with high risk-aversion 0.9, to learn to minimize the worst-case discounted cost with $\gamma = 0.97$. We compare the worst-case performance of the greedy strategy of the trained AIS-DQN with the worst-case performance of a trained stochastic-DQN, which uses the observation and position as the state and has the same hyperparameters with no risk aversion. In Fig. 2, we present the improvement in worst-case cost achieved by AIS-DQN over stochastic-DQN in 10^4 simulations each for different initial positions. Note that AIS-DQN outperforms stochastic-DQN for most cases.

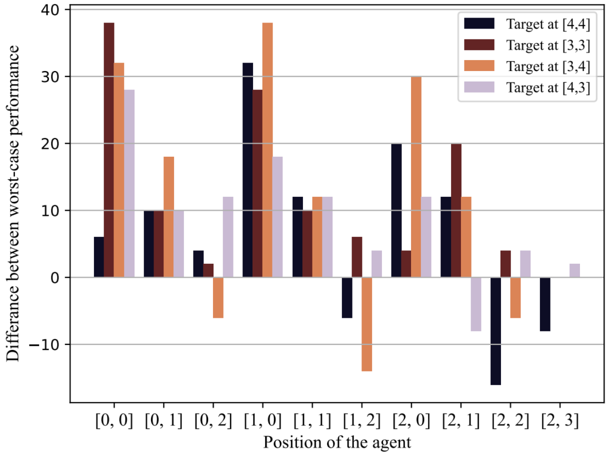


Fig. 2. The improvement in worst-case performance using AIS-DQL over stochastic-DQL.

VI. Conclusions

In this paper, we provided a general notion of information states for worst-case decision-making problems over an infinite time horizon with observable costs. We showed that these information states yield a time-invariant DP decomposition to compute an optimal control strategy. Then, we extended this notion to define approximate information states and an approximate DP. We proved the associated approximation bounds in worst-case performance. Finally, we illustrated, using a numerical example, how approximate information states can be learned using output data and used to generate control strategies. Future work should consider using these results in applications requiring approximately worst-case control and worst-case reinforcement learning.

References

[1] A. A. Malikopoulos, L. E. Beaver, and I. V. Chremos, "Optimal time trajectory and coordination for connected and automated vehicles," *Automatica*, vol. 125, no. 109469, 2021.

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

[3] A. Dave, N. Venkatesh, and A. A. Malikopoulos, "On decentralized control of two agents with nested accessible information," in *2022 American Control Conference (ACC)*, pp. 3423–3430, IEEE, 2022.

[4] S. Mannor, D. Simester, P. Sun, and J. N. Tsitsiklis, "Bias and variance approximation in value function estimates," *Management Science*, vol. 53, no. 2, pp. 308–322, 2007.

[5] W. Wiesemann, D. Kuhn, and B. Rustem, "Robust markov decision processes," *Mathematics of Operations Research*, vol. 38, no. 1, pp. 153–183, 2013.

[6] O. Mihatsch and R. Neuneier, "Risk-sensitive reinforcement learning," *Machine learning*, vol. 49, pp. 267–290, 2002.

[7] N. Bäauerle and U. Rieder, "Partially observable risk-sensitive markov decision processes," *Mathematics of Operations Research*, vol. 42, no. 4, pp. 1180–1196, 2017.

[8] M. Rasouli, E. Miehling, and D. Teneketzis, "A scalable decomposition method for the dynamic defense of cyber networks," in *Game Theory for Security and Risk Management*, pp. 75–98, Springer, 2018.

[9] Y. Shoukry, J. Araujo, P. Tabuada, M. Srivastava, and K. H. Johansson, "Minimax control for cyber-physical systems under network packet scheduling attacks," in *Proceedings of the 2nd ACM international conference on High confidence networked systems*, pp. 93–100, 2013.

[10] D. Bertsekas and I. Rhodes, "Sufficiently informative functions and the minimax feedback control of uncertain dynamic systems," *IEEE Transactions on Automatic Control*, vol. 18, no. 2, pp. 117–124, 1973.

[11] P. Bernhard, "Max-plus algebra and mathematical fear in dynamic optimization," *Set-Valued Analysis*, vol. 8, no. 1, pp. 71–84, 2000.

[12] M. Gagrani and A. Nayyar, "Decentralized minimax control problems with partial history sharing," in *2017 American Control Conference (ACC)*, pp. 3373–3379, IEEE, 2017.

[13] S. P. Coraluppi and S. I. Marcus, "Risk-sensitive and minimax control of discrete-time, finite-state markov decision processes," *Automatica*, vol. 35, no. 2, pp. 301–309, 1999.

[14] A. Dave, N. Venkatesh, and A. A. Malikopoulos, "On decentralized minimax control with nested subsystems," in *2022 American Control Conference (ACC)*, pp. 3437–3444, IEEE, 2022.

[15] A. Dave, N. Venkatesh, and A. A. Malikopoulos, "Approximate Information States for Worst-Case Control and Learning in Uncertain Systems," *arXiv:2301.05089 (in review)*, 2023.

[16] M. R. James, J. S. Baras, and R. J. Elliott, "Risk-sensitive control and dynamic games for partially observed discrete-time nonlinear systems," *IEEE transactions on automatic control*, vol. 39, no. 4, pp. 780–792, 1994.

[17] P. Bernhard, "Minimax - or feared value - L_1 / L_∞ control," *Theoretical computer science*, vol. 293, no. 1, pp. 25–44, 2003.

[18] J. S. Baras and N. S. Patel, "Robust control of set-valued discrete-time dynamical systems," *IEEE Transactions on Automatic Control*, vol. 43, no. 1, pp. 61–75, 1998.

[19] A. Dave, N. Venkatesh, and A. A. Malikopoulos, "Approximate information states for worst-case control of uncertain systems," in *Proceedings of the 61th IEEE Conference on Decision and Control (CDC)*, pp. 4945–4950, 2022.

[20] A. Dave, N. Venkatesh, and A. A. Malikopoulos, "On robust control of partially observed uncertain systems with additive costs," in *Proceedings of the 2023 American Control Conference (ACC)*, pp. 4639–4644, 2023.

[21] J. Subramanian, A. Sinha, R. Seraj, and A. Mahajan, "Approximate information state for approximate planning and reinforcement learning in partially observed systems," *Journal of Machine Learning Research*, vol. 23, no. 12, pp. 1–83, 2022.

[22] G. N. Nair, "A nonstochastic information theory for communication and state estimation," *IEEE Transactions on automatic control*, vol. 58, no. 6, pp. 1497–1510, 2013.

[23] A. Dave, I. Faros, N. Venkatesh, and A. A. Malikopoulos, "Worst-case control and learning using partial observations over an infinite time-horizon," *arXiv preprint, arXiv: 2303.16321*, 2023.

[24] D. Karimi and S. E. Salcudean, "Reducing the hausdorff distance in medical image segmentation with convolutional neural networks," *IEEE Transactions on medical imaging*, vol. 39, no. 2, pp. 499–513, 2019.