

Fundamental Performance Limits for Sensor-Based Robot Control and Policy Learning

Anirudha Majumdar and Vincent Pacelli
Department of Mechanical and Aerospace Engineering
Princeton University, Princeton, NJ, 08540
Emails: {ani.majumdar, vpacelli}@princeton.edu

Abstract—Our goal is to develop theory and algorithms for establishing *fundamental limits* on performance for a given task imposed by a robot’s sensors. In order to achieve this, we define a quantity that captures the amount of *task-relevant information* provided by a sensor. Using a novel version of the generalized Fano inequality from information theory, we demonstrate that this quantity provides an upper bound on the highest achievable expected reward for one-step decision making tasks. We then extend this bound to multi-step problems via a dynamic programming approach. We present algorithms for numerically computing the resulting bounds, and demonstrate our approach on three examples: (i) the lava problem from the literature on partially observable Markov decision processes, (ii) an example with continuous state and observation spaces corresponding to a robot catching a freely-falling object, and (iii) obstacle avoidance using a depth sensor with non-Gaussian noise. We demonstrate the ability of our approach to establish strong limits on achievable performance for these problems by comparing our upper bounds with achievable lower bounds (computed by synthesizing or learning concrete control policies).

I. INTRODUCTION

Robotics is often characterized as the problem of transforming “pixels to torques” [1]: how can an embodied agent convert raw sensor inputs into actions in order to accomplish a given task? In this paper, we seek to understand the *fundamental limits* of this process by studying the following question: is there an *upper bound* on performance imposed by the sensors that a robot is equipped with?

As a motivating example, consider the recent debate around the “camera-only” approach to autonomous driving favored by Tesla versus the “sensor-rich” philosophy pursued by Waymo [2]. Is an autonomous vehicle equipped only with cameras *fundamentally limited* in terms of the performance or safety it can achieve? By “fundamental limit”, we mean a bound on performance (or safety) on a given task that holds *regardless* of the form of control policy one utilizes (e.g., a neural network with billions of parameters, a nonlinear model predictive control scheme combined with a particle filter, etc.), how the policy is synthesized (e.g., via model-free reinforcement learning, model-based control, etc.), or how much computation is available to the robot or software designer.

While there have been tremendous algorithmic advancements in robotics over decades, we currently lack a “science” for understanding such fundamental limits [3]. Current practice in robotics is often *empirical* in nature (e.g., trying different perception and control architectures with neural networks

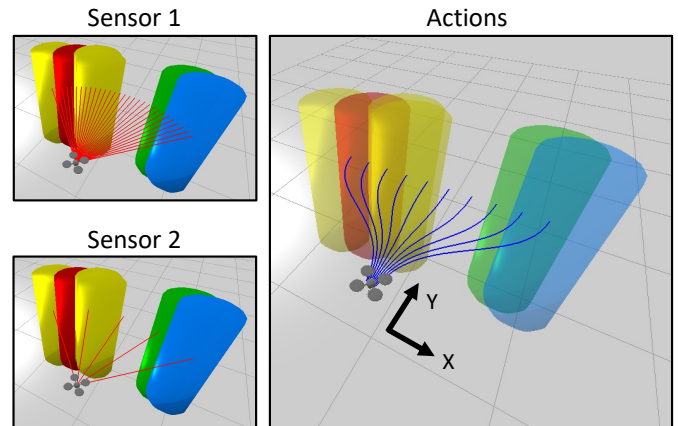


Fig. 1. Our goal is to establish fundamental limits on performance for a given task imposed by a robot’s sensors. We demonstrate our approach on examples that include obstacle avoidance with a noisy depth sensor (fig. left) using motion primitives (fig. right). We also show how we can use our approach to establish the superiority of one sensor (Sensor 1: a dense depth sensor) over another (Sensor 2: a sparse depth sensor) for a given task.

of growing size and varying architectures). Techniques for establishing fundamental limits imposed by a sensor would potentially allow us to glean important design insights (e.g., realizing that a particular sensor is not sufficient for a task and must be replaced). Further, such techniques could allow us to establish the superiority of one suite of sensors over another from the perspective of a given task, e.g., by synthesizing a control policy for one sensor suite that achieves better performance than the fundamental bound for another suite.

In this paper, we take a step towards this goal. We first observe that any technique for establishing fundamental bounds on performance imposed by a given sensor must take into account two factors: (i) the *quality* of the sensor (e.g., the amount of information about the state of the robot and its environment provided by the sensor), and, importantly, (ii) the *task* that the robot is meant to accomplish. As an example, consider a drone equipped with a (noisy) depth sensor (Fig. 1). Depending on the nature of the task, the robot may need more or less information from its sensors. For example, suppose that the obstacle locations are highly constrained such that a particular sequence of actions always succeeds in avoiding them (i.e., there is a purely open-loop policy that achieves good performance on the task); in this case, even an extremely noisy or sparse depth sensor allows the robot to perform well. However, if the distribution of obstacles is such that there is no

pre-defined gap in the obstacles, then a noisy or sparse depth sensor may fundamentally limit the achievable performance on the task. The achievable performance is thus intuitively influenced by the amount of *task-relevant information* provided by the robot's sensors.

Statement of contributions. Our primary contribution is to develop theory and algorithms for establishing fundamental bounds on performance imposed by a robot's sensors for a given task. Our key insight is to define a quantity that captures the *task-relevant information* provided by the robot's sensors. Using a novel version of the *generalized Fano inequality* from information theory, we demonstrate that this quantity provides a fundamental upper bound on expected reward for one-step decision making problems. We then extend this bound to multi-step settings via a dynamic programming approach and propose algorithms for computing the resulting bounds for systems with potentially continuous state and observation spaces, nonlinear and stochastic dynamics, and non-Gaussian sensor models (but with discretized action spaces). We demonstrate our approach on three examples: (i) the lava problem from the literature on partially observable Markov decision processes (POMDPs), (ii) a robot catching a freely-falling object, and (iii) obstacle avoidance using a depth sensor (Fig. 1). We demonstrate the strength of our bounds on these examples by comparing them against the performance achieved by concrete control policies: the optimal POMDP solution for the lava problem, a model-predictive control (MPC) scheme for the catching example, and a learned neural network policy for the obstacle avoidance problem. We also present applications of our approach for establishing the superiority of one sensor over another (from the perspective of a given task). To our knowledge, the results in this paper are the first to provide general-purpose techniques for establishing fundamental bounds on performance for sensor-based control of robots.

A. Related Work

Domain-specific performance bounds. Prior work in robotics has established fundamental bounds on performance for particular problems. For example, [4, 5] consider high-speed navigation through an (ergodic) forest consisting of randomly-placed obstacles. Results from percolation theory [6] are used to establish a critical speed beyond which there does not exist (with probability one) an infinite collision-free trajectory. The work in [7] establishes limits on the speed at which a robot can navigate through unknown environments in terms of perceptual latency. Classical techniques from robust control [8] have also been utilized to establish fundamental limits on performance for control tasks (e.g., pole balancing) involving linear output-feedback control and sensor noise or delays [9]. The results in [10] demonstrate empirical correlation of the complexity metrics presented in [9] with sample-efficiency and performance of learned perception-based controllers on a pole-balancing task. The approaches mentioned above consider specific tasks (e.g., navigation in ergodic forests) or relatively narrow classes of problems (e.g., linear output-feedback control). In contrast, our goal is to develop a general

and broadly applicable theoretical and algorithmic framework for establishing fundamental bounds on performance imposed by a sensor for a given task.

Comparing sensors. The notion of a *sensor lattice* was introduced in [11, 12] for comparing the power of different sensors (see [13] for a similar approach for comparing robots). The sensor lattice provides a partial ordering on different sensors based on the ability of one sensor to simulate another. However, most pairs of sensors are *incomparable* using such a scheme. Moreover, the sensor lattice does not establish the superiority of one sensor over another from the perspective of a given task; instead, the partial ordering is based on the ability of one sensor to perform as well as another in terms of filtering (i.e., state estimation). In this paper, we also demonstrate the applicability of our approach for comparing different sensors. However, this comparison is *task-driven*; we demonstrate how one sensor can be proved to be fundamentally better than another from the perspective of a given task.

Fano's inequality and its extensions. In its original form, *Fano's inequality* [14] relates the lowest achievable error of estimating a signal x from an observation y in terms of the noise in the channel that produces observations from signals. In recent years, Fano's inequality has been significantly extended and applied for establishing fundamental limits for various statistical estimation problems, e.g., lower bounding the Bayes and minimax risks for different learning problems [15–17]. In this paper, we build on generalized versions of Fano's inequality [16, 17] in order to obtain fundamental bounds on performance for robotic systems with noisy sensors. On the technical front, we contribute by deriving a stronger version of the generalized Fano inequalities presented in [16, 17] by utilizing the *inverse of the KL divergence* (Sec. III) and computing it using *geometric programming* [18, Ch. 4.5]. The resulting inequality, which may be of independent interest, allows us to derive fundamental upper bounds on performance for one-step decision making problems. We then develop a dynamic programming approach for recursively applying the generalized Fano inequality in order to derive bounds on performance for multi-step problems.

II. PROBLEM FORMULATION

A. Notation

We denote sequences by $x_{i:j} := (x_k)_{k=i}^j$ for $i \leq j$. Expectations are typically denoted as $\mathbb{E}[\cdot]$ with the variable of integration or its measure appearing below it for contextual emphasis, e.g.: $\mathbb{E}_x[\cdot]$, $\mathbb{E}_{p(x)}[\cdot]$.

B. Problem Statement

We denote the state of the robot and its environment at time-step t by $s_t \in \mathcal{S}$. Let p_0 denote the initial state distribution. Let the robot's sensor observation and control action at time-step t be denoted by $o_t \in \mathcal{O}$ and $a_t \in \mathcal{A}$ respectively. Denote the (stochastic) dynamics of the state by $p_t(s_t|s_{t-1}, a_{t-1})$ and suppose that the robot's sensor is described by $\sigma_t(o_t|s_t)$. The robot's task is prescribed using reward functions $r_0, r_1, \dots, r_{T-1} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ at each time-step (up to a finite horizon).

Assumption 1 (Bounded rewards). *We assume that rewards are bounded, and without further loss of generality we assume that $r_t(s_t, a_t) \in [0, 1]$, $\forall s_t \in \mathcal{S}, a_t \in \mathcal{A}, t \in \{0, \dots, T-1\}$.*

The robot's goal is to find a (potentially time-varying and history-dependent) control policy $\pi_t : \mathcal{O}^{t+1} \rightarrow \mathcal{A}$ that maximizes the total expected reward:

$$R^* := \sup_{\pi_{0:T-1}} \mathbb{E}_{\substack{s_{0:T-1} \\ o_{0:T-1}}} \left[\sum_{t=0}^{T-1} r_t(s_t, \pi_t(o_{0:t})) \right]. \quad (1)$$

Goal: Our goal is to *upper bound* the best achievable expected reward R^* for a given sensor $\sigma_{0:T-1}$. We note that we are allowing for completely general policies that are arbitrary time-varying functions of the entire history of observations received up to time t (as long as the functions satisfy measurability conditions that ensure the existence of the expectation in (1)). An upper bound on R^* thus provides a fundamental bound on achievable performance that holds regardless of how the policy is parameterized (e.g., via neural networks, receding-horizon control architectures, etc.) or synthesized (e.g., via reinforcement learning, optimal control techniques, etc.).

III. BACKGROUND

In this section, we briefly introduce some background material that will be useful throughout the paper.

A. KL Divergence and Mutual Information

The Kullback-Leibler (KL) divergence between two distributions is defined as:

$$\mathbb{D}(p(x) \| q(x)) := \mathbb{E}_{p(x)} \left[\log \frac{p(x)}{q(x)} \right]. \quad (2)$$

The mutual information between two random variables x and y is defined as:

$$\mathbb{I}(x; y) := \mathbb{D}(p(x, y) \| p(x)p(y)), \quad (3)$$

where $p(x, y)$ is the joint distribution, and $p(x)$ and $p(y)$ are the resulting marginal distributions. This quantity captures the amount of information one obtains about one random variable (e.g., the state s_t) by observing another random variable (e.g., sensor observations o_t).

B. Inverting Bounds on the KL Divergence

Let $\mathcal{B}_p$ and $\mathcal{B}_q$ be Bernoulli distributions on $\{0, 1\}$ with mean p and q respectively. For $p, q \in [0, 1]$, we define¹:

$$\mathbb{D}_{\mathcal{B}}(p \| q) := \mathbb{D}(\mathcal{B}_p \| \mathcal{B}_q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}.$$

In subsequent sections, we will obtain bounds on a quantity $p^* \in [0, 1]$ given by $\mathbb{D}_{\mathcal{B}}(p^* \| q) \leq c$ for some $q \in [0, 1]$ and $c \geq 0$. In order to upper bound p^* , we will use the *KL inverse*:

$$\mathbb{D}^{-1}(q|c) := \sup \{p \in [0, 1] \mid \mathbb{D}_{\mathcal{B}}(p \| q) \leq c\}. \quad (4)$$

¹We adopt the commonly used convention when working with the KL divergence of taking $0 \cdot \log(0/a) := 0, \forall a$, and $a \cdot \log(a/0) := \infty, \forall a > 0$.

It is then easy to see that $p^* \leq \mathbb{D}^{-1}(q|c)$.

Since $\mathbb{D}_{\mathcal{B}}(\cdot \| \cdot)$ is (jointly) convex in both arguments, the optimization problem in (4) is a convex problem. In particular, one can compute the KL inverse efficiently using a *geometric program* [18, Ch. 4.5] with a single decision variable p .

IV. PERFORMANCE BOUND FOR SINGLE-STEP PROBLEMS

In this section, we will derive an upper bound on the best achievable reward R^* in the single time-step decision-making setting. This bound will then be extended to the multi-step setting in Sec. V.

When $T = 1$, our goal is to upper bound the following quantity:

$$R^*(\sigma_0; r_0) := \sup_{\pi_0} \mathbb{E}_{s_0, o_0} [r_0(s_0, \pi_0(o_0))] \quad (5)$$

$$= \sup_{\pi_0} \mathbb{E}_{p_0(s_0)} \mathbb{E}_{\sigma_0(o_0|s_0)} [r_0(s_0, \pi_0(o_0))]. \quad (6)$$

The notation $R^*(\sigma_0; r_0)$ highlights the dependence of the best achievable reward in terms of the robot's sensor and task (as specified by the reward function). As highlighted in Sec. I, the amount of information that the robot requires from its sensors in order to obtain high expected reward depends on its task; certain tasks may admit purely open-loop policies that obtain high rewards, while other tasks may require high-precision sensing of the state. We formally define a quantity that captures this intuition and quantifies the *task-relevant information* provided by the robot's sensors. We then demonstrate that this quantity provides an upper bound on $R^*(\sigma_0; r_0)$.

Definition 1 (Task-relevant information potential). *Let $\mathbb{I}(o_0; s_0)$ be the mutual information between the robot's sensor observation and state. Define:*

$$R_0^\perp := \sup_{a_0} \mathbb{E}_{s_0} [r_0(s_0, a_0)] \quad (7)$$

as the highest achievable reward using an open-loop policy. Then define the task-relevant information potential (TRIP) of a sensor σ_0 for a task specified by reward function r_0 as:

$$\tau(\sigma_0; r_0) := \mathbb{D}^{-1}(R_0^\perp | \mathbb{I}(o_0; s_0)). \quad (8)$$

In order to interpret the TRIP, we state two useful properties of the KL inverse.

Proposition 1 (Monotonicity of KL inverse). *The KL inverse $\mathbb{D}^{-1}(q|c)$ is:*

- 1) *monotonically non-decreasing in $c \geq 0$ for fixed $q \in [0, 1]$,*
- 2) *monotonically non-decreasing in $q \in [0, 1]$ for fixed $c \geq 0$.*

Proof: The first property follows from the fact that increasing c loosens the KL divergence constraint in the optimization problem in (4). The proof of the second property utilizes the envelope theorem [19, Corollary 5] and is provided in Appendix A (Lemma 4). ■

The TRIP $\tau(\sigma_0; r_0)$ depends on two factors: the mutual information $\mathbb{I}(o_0; s_0)$ (which depends on the robot's sensor) and

the best reward $R_0^\perp$ achievable by an open-loop policy (which depends on the robot's task). Using Proposition 1, we see that as the sensor provides more information about the state (i.e., as $\mathbb{I}(o_0; s_0)$ increases for fixed $R_0^\perp$), the TRIP is monotonically non-decreasing. Moreover, the TRIP is a monotonically non-decreasing function of $R_0^\perp$ for fixed $\mathbb{I}(o_0; s_0)$. This qualitative dependence is intuitively appealing: if there is a good open-loop policy (i.e., one that achieves high reward), then the robot's sensor can provide a small amount of information about the state and still lead to good overall performance. The specific form of the definition of TRIP is motivated by the result below, which demonstrates that the TRIP upper bounds the best achievable expected reward $R^*(\sigma_0; r_0)$ in Eq. (5).

Theorem 1 (Single-step performance bound). *The best achievable reward is upper bounded by the task-relevant information potential (TRIP) of a sensor:*

$$\tau(\sigma_0; r_0) \geq R^*(\sigma_0; r_0) := \sup_{\pi_0} \mathbb{E}_{s_0, o_0} [r_0(s_0, \pi_0(o_0))]. \quad (9)$$

Proof: The proof is provided in Appendix A and is inspired by the proof of the generalized Fano inequality presented in [17, Proposition 14]. The bound (9) tightens the generalized Fano inequality [16, 17] by utilizing the KL inverse (in contrast to the methods in [16, 17], which may be interpreted as indirectly bounding the KL inverse). The result presented here may thus be of independent interest. ■

Theorem 1 provides a *fundamental bound* on performance (in the sense of Sec. I) imposed by the sensor for a given single-step task. This bound holds for *any* policy, independent of its complexity or how it is synthesized or learned.

V. PERFORMANCE BOUND FOR MULTI-STEP PROBLEMS: FANO'S INEQUALITY WITH FEEDBACK

In this section, we derive an upper bound on the best achievable reward R^* defined in (1) for the general multi time-step setting. The key idea is to extend the single-step bound from Theorem 1 using a dynamic programming argument.

Let $\pi_k^t : \mathcal{O}^{k-t+1} \rightarrow \mathcal{A}$ denote a policy that takes as input the sequence of observations $o_{t:k}$ from time-step t to k (for $k \geq t$). Thus, a policy π_k^0 at time-step k utilizes all observations received up to time-step k . Given an initial state distribution p_0 and an open-loop action sequence $a_{0:t-1}$, define the reward-to-go from time $t \in \{0, \dots, T-1\}$ given $a_{0:t-1}$ as:

$$R_t := \mathbb{E}_{\substack{s_{t:T-1}, o_{t:T-1} \\ a_{0:t-1}}} \left[\sum_{k=t}^{T-1} r_k(s_k, \pi_k^t(o_{t:k})) \right], \quad (10)$$

where the expectation,

$$\mathbb{E}_{\substack{s_{t:T-1}, o_{t:T-1} \\ a_{0:t-1}}} \left[\cdot \right] \quad (11)$$

is taken with respect to the distribution of states $s_{t:T-1}$ and observations $o_{t:T-1}$ one receives if one propagates p_0 using

the open-loop sequence of actions from time-steps² 0 to $t-1$, and then applies the closed-loop policies $\pi_t^t, \pi_{t+1}^t, \dots, \pi_{T-1}^t$ from time-steps t to $T-1$. We further define $R_T := 0$.

Now, for $t \in \{0, \dots, T-1\}$, define:

$$R_t^\perp := \sup_{a_t} \left[\mathbb{E}_{s_t | a_{0:t-1}} [r_t(s_t, a_t)] + R_{t+1} \right], \quad (13)$$

and

$$R_t^{\perp*} := \sup_{\pi_{t+1}^{t+1}, \dots, \pi_{T-1}^{t+1}} R_t^\perp. \quad (14)$$

The following result then leads to a recursive structure for computing an upper bound on R^* .

Proposition 2 (Recursive bound). *For any $t = 0, \dots, T-1$, the following inequality holds for any open-loop sequence of actions $a_{0:t-1}$:*

$$\sup_{\pi_t^t, \dots, \pi_{T-1}^t} R_t \leq \underbrace{(T-t) \cdot \mathbb{D}^{-1} \left(\frac{R_t^{\perp*}}{T-t} \mid \mathbb{I}(o_t; s_t) \right)}_{=:\tau_t(\sigma_{t:T-1}; r_{t:T-1})}. \quad (15)$$

Proof: The proof follows a similar structure to Theorem 1 and is presented in Appendix A. ■

To see how we can use Proposition 2, we first use (1) and (10) to note that the LHS of (15) for $t = 0$ is equal to R^* :

$$R^* = \sup_{\pi_0^0, \dots, \pi_{T-1}^0} R_0 \leq \tau_0(\sigma_{0:T-1}; r_{0:T-1}). \quad (16)$$

The quantity $\tau_t(\sigma_{t:T-1}; r_{t:T-1})$ may be interpreted as a multi-step version of the TRIP from Definition 1. This quantity depends on the mutual information $\mathbb{I}(o_t; s_t)$, which is computed using the distribution $p_t(s_t | a_{0:t-1})$ over s_t that one obtains by propagating p_0 using the open-loop sequence of actions $a_{0:t-1}$:

$$\mathbb{I}(o_t; s_t) = \mathbb{D} \left(p_t(s_t | a_{0:t-1}) \sigma_t(o_t | s_t) \parallel p_t(s_t | a_{0:t-1}) \sigma_t(o_t) \right). \quad (17)$$

In addition, $\tau_t(\sigma_{t:T-1}; r_{t:T-1})$ depends on $R_t^{\perp*}$, which is then divided by $(T-t)$ to ensure boundedness between $[0, 1]$ (see Assumption 1). The quantity $R_t^{\perp*}$ can itself be upper bounded using (15) with $t+1$, as we demonstrate below. Such an upper bound on $R_t^{\perp*}$ for $t = 0$ leads to an upper bound on R^* using (16) and the monotonicity of the KL inverse (Proposition 1). Applying this argument recursively leads to Algorithm 1, which computes an upper bound on R^* .

²For $t = 0$, we use the convention that $a_{0:-1}$ is the empty sequence and:

$$\mathbb{E}_{\substack{s_{0:T-1}, o_{0:T-1} \\ a_{0:-1}}} \left[\cdot \right] := \mathbb{E}_{s_{0:T-1}, o_{0:T-1}} \left[\cdot \right]. \quad (12)$$

Algorithm 1 Multi-Step Performance Bound

```

1: Initialize  $\bar{R}_T(a_{0:T-1}) = 0, \forall a_{0:T-1}$ .
2: for  $t = T-1, T-2, \dots, 0$  do
3:    $\bar{R}_t(a_{0:t-1}) = (T-t) \cdot \mathbb{D}^{-1}\left(\frac{\bar{R}_t^{\perp*}}{T-t} \mid \mathbb{I}(o_t; s_t)\right), \forall a_{0:t-1}$ ,
     where  $\bar{R}_t^{\perp*} := \sup_{a_t} \mathbb{E}_{s_t|a_{0:t-1}} \left[ r_t(s_t, a_t) \right] + \bar{R}_{t+1}(a_{0:t})$ .
4: end for
5: return  $\bar{R}_0$  (bound on achievable expected reward).

```

Theorem 2 (Multi-step performance bound). *Algorithm 1 returns an upper bound on the best achievable reward R^* .*

Proof: We provide a sketch of the proof here, which uses (backwards) induction. In particular, Proposition 2 leads to the inductive step. See Appendix A for the complete proof.

We prove that for all $t = T-1, \dots, 0$,

$$\sup_{\pi_t^t, \dots, \pi_{T-1}^t} R_t \leq \bar{R}_t(a_{0:t-1}), \quad \forall a_{0:t-1}. \quad (18)$$

Thus, in particular,

$$R^* = \sup_{\pi_0^0, \dots, \pi_{T-1}^0} R_0 \leq \bar{R}_0. \quad (19)$$

We prove (18) by backwards induction starting from $t = T-1$. We first prove the base step. Using (15), we obtain:

$$\sup_{\pi_{T-1}^{T-1}} R_{T-1} \leq \mathbb{D}^{-1}\left(\bar{R}_{T-1}^{\perp*} \mid \mathbb{I}(o_{T-1}; s_{T-1})\right). \quad (20)$$

Using the fact that $R_T = 0$, we can show that $R_{T-1}^{\perp*} = \bar{R}_{T-1}^{\perp*}$. Combining this with (20) and the monotonicity of the KL inverse (Proposition 1), we see:

$$\sup_{\pi_{T-1}^{T-1}} R_{T-1} \leq \mathbb{D}^{-1}\left(\bar{R}_{T-1}^{\perp*} \mid \mathbb{I}(o_{T-1}; s_{T-1})\right) \quad (21)$$

$$= \bar{R}_{T-1}(a_{0:T-2}). \quad (22)$$

In order to prove the induction step, suppose that for $t \in \{0, \dots, T-2\}$, we have

$$\sup_{\pi_{t+1}^{t+1}, \dots, \pi_{T-1}^{t+1}} R_{t+1} \leq \bar{R}_{t+1}(a_{0:t}). \quad (23)$$

We then need to show that

$$\sup_{\pi_t^t, \dots, \pi_{T-1}^t} R_t \leq \bar{R}_t(a_{0:t-1}). \quad (24)$$

We can use the induction hypothesis (23) to show that $R_t^{\perp*} \leq \bar{R}_t^{\perp*}$. Combining this with (15) and the monotonicity of the KL inverse (Proposition 1), we obtain the desired result (24):

$$\sup_{\pi_t^t, \dots, \pi_{T-1}^t} R_t \leq (T-t) \cdot \mathbb{D}^{-1}\left(\frac{\bar{R}_t^{\perp*}}{T-t} \mid \mathbb{I}(o_t; s_t)\right) \quad (25)$$

$$= \bar{R}_t(a_{0:t-1}). \quad (26)$$

■

VI. NUMERICAL IMPLEMENTATION

In order to compute the single-step bound using Theorem 1 or the multi-step bound using Algorithm 1, we require the ability to compute (or bound) three quantities: (i) the KL inverse, (ii) the mutual information $\mathbb{I}(o_t; s_t)$, and (iii) the quantity $\bar{R}_t^{\perp*}$. As described in Sec. III-B, we can compute the KL inverse efficiently using a geometric program (GP) [18, Ch. 4.5] with a single decision variable. There are a multitude of solvers for GPs including Mosek [20] and the open-source solver SCS [21]. Next, we describe the computation of $\mathbb{I}(o_t; s_t)$ and $\bar{R}_t^{\perp*}$ in different settings.

A. Analytic Computation

In certain settings, one can compute $\mathbb{I}(o_t; s_t)$ and $\bar{R}_t^{\perp*}$ exactly. We discuss two such settings of interest below.

Discrete POMDPs. In cases where the state space $\mathcal{S}$, action space $\mathcal{A}$, and observation space $\mathcal{O}$ are finite, one can compute $\mathbb{I}(o_t; s_t)$ exactly by propagating the initial state distribution p_0 forward using open-loop action sequences $a_{0:t-1}$ and using the expression (17) (which can be evaluated exactly since we have discrete probability distributions). The expectation term in $\bar{R}_t^{\perp*}$ can be computed similarly. In addition, the supremum over actions can be evaluated exactly via enumeration.

Linear-Gaussian systems with finite action spaces. One can also perform exact computations in cases where (i) the state space $\mathcal{S}$ is continuous and the dynamics $p_t(s_t|s_{t-1}, a_{t-1})$ are given by a linear dynamical system with additive Gaussian uncertainty, (ii) the observation space $\mathcal{O}$ is continuous and the sensor model $\sigma_t(o_t|s_t)$ is such that the observations are linear (in the state) with additive Gaussian uncertainty, (iii) the initial state distribution p_0 is Gaussian, and (iv) the action space $\mathcal{A}$ is finite. In such settings, one can analytically propagate p_0 forward through open-loop action sequences $a_{0:t-1}$ using the fact that Gaussian distributions are preserved when propagated through linear-Gaussian systems (similar to Kalman filtering [22]). One can then compute $\mathbb{I}(o_t; s_t)$ using (17) by leveraging the fact that all the distributions involved are Gaussian, for which KL divergences can be computed in closed form [23]. One can also compute $\bar{R}_t^{\perp*}$ exactly for any reward function that permits the analytic computation of the expectation term using a Gaussian (e.g., quadratic reward functions); the supremum over actions can be evaluated exactly since $\mathcal{A}$ is finite.

B. Computation via Sampling and Concentration Inequalities

General settings. Next, we consider more general settings with: (i) continuous state and observation spaces, (ii) arbitrary (e.g, non-Gaussian/nonlinear) dynamics $p_t(s_t|s_{t-1}, a_{t-1})$, which are potentially not known analytically, but can be sampled from (e.g., as in a simulator), (iii) arbitrary (e.g., non-Gaussian/nonlinear) sensor $\sigma_t(o_t|s_t)$, but with a probability density function that can be numerically evaluated given any particular state-observation pair, (iv) an arbitrary initial state distribution p_0 that can be sampled from, and (v) a finite action space. Our bound is thus broadly applicable, with the primary restriction being the finiteness of $\mathcal{A}$; we leave extensions to continuous action spaces for future work (see Sec. VIII).

We first discuss the computation of $\bar{R}_t^{\perp*}$. Since the supremization over actions can be performed exactly (due to the finiteness of $\mathcal{A}$), the primary challenge here is to evaluate the expectation:

$$\mathbb{E}_{s_t|a_{0:t-1}} \left[r_t(s_t, a_t) \right]. \quad (27)$$

We note that any upper bound on this expectation leads to an upper bound on $\bar{R}_t^{\perp*}$, and thus a valid upper bound on R^* (due to the monotonicity of the KL inverse; Proposition 1). One can thus obtain a high-confidence upper bound on (27) by sampling states $s_t|a_{0:t-1}$, and using any concentration inequality [24]. In particular, since we assume boundedness of rewards (Assumption 1), we can use Hoeffding’s inequality.

Theorem 3 (Hoeffding’s inequality [24]). *Let z be a random variable bounded within $[0, 1]$, and let $z_1, \dots, z_n$ denote i.i.d. samples. Then, with probability at least $1 - \delta$ (over the sampling of $z_1, \dots, z_n$), the following bound holds with probability at least $1 - \delta$:*

$$\mathbb{E}[z] \leq \frac{1}{n} \sum_{i=1}^n z_i + \sqrt{\frac{\log(1/\delta)}{2n}}. \quad (28)$$

In our numerical examples (Sec. VII), we utilize a slightly tighter version of Hoeffding’s inequality (see Appendix B).

Next, we discuss the computation of $\mathbb{I}(o_t; s_t)$. Again, we note that any upper bound on $\mathbb{I}(o_t; s_t)$ yields a valid upper bound on R^* due to the monotonicity of the KL inverse. In this work, we utilize *variational bounds* on mutual information; in particular, we use the “leave-one-out bound” [25]:

$$\mathbb{I}(o_t; s_t) \leq \mathbb{E} \left[\frac{1}{K} \sum_{i=1}^K \left[\log \frac{\sigma_t(o_t^{[i]} | s_t^{[i]})}{\frac{1}{K-1} \sum_{j \neq i} \sigma_t(o_t^{[i]} | s_t^{[j]})} \right] \right], \quad (29)$$

where the expectation is over size- K batches $\{(s_t^{[i]}, o_t^{[i]})\}_{i=1}^K$ of sampled states $s_t|a_{0:t-1}$ and observations sampled using $\sigma_t(o_t | s_t)$. The quantity $\sigma_t(o_t^{[i]} | s_t^{[i]})$ denotes (with slight abuse of notation) the evaluation of the density function corresponding to the sensor model. Since the bound (29) is in terms of an expectation, one can again obtain a high-confidence upper bound by sampling state-observation batches and applying a concentration inequality (e.g., Hoeffding’s inequality if the quantity inside the expectation is bounded).

We note that the overall implementation of Algorithm 1 may involve the application of multiple concentration inequalities (each of which holds with some confidence $1 - \delta_i$). One can obtain the overall confidence of the upper bound on R^* by using a union bound: $1 - \delta = 1 - \sum_i \delta_i$.

C. Tightening the Bound with Multiple Horizons

We end this section by discussing a final implementation detail. Let T denote the time horizon of interest (as in Sec. II). For any $H \in \{1, \dots, T\}$, one can define

$$R_H^* := \sup_{\pi_{0:H-1}} \mathbb{E}_{s_0:H-1, o_0:H-1} \left[\sum_{t=0}^{H-1} r_t(s_t, \pi_t(o_{0:t})) \right] \quad (30)$$

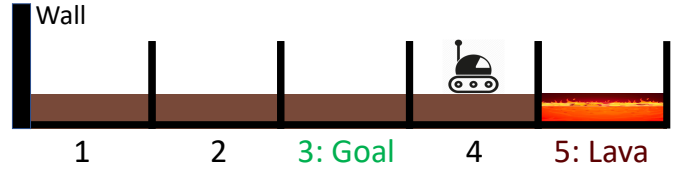


Fig. 2. An illustration of the lava problem. The robot needs to navigate to a goal without falling into the lava (using a noisy sensor).

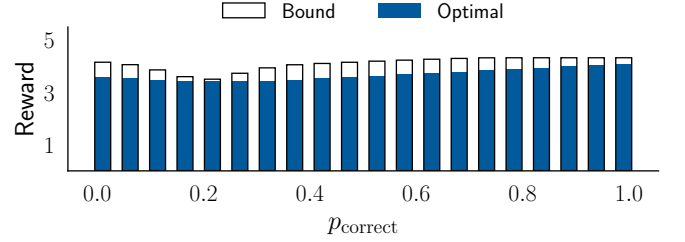


Fig. 3. Results for the lava problem. We compare the upper bounds on achievable expected rewards computed by our approach with the optimal POMDP solution for different values of sensor noise.

as the best achievable reward for a problem with horizon H (instead of T). One can then apply Algorithm 1 to compute an upper bound on R_H^* . Since rewards are assumed to be bounded within $[0, 1]$ (Assumption 1), we can observe that $R^* = R_T^* \leq R_H^* + (T - H)$. In practice, we sometimes find that this bound provides a tighter bound on R^* for some $H < T$ (as compared to directly applying Algorithm 1 with a horizon of T). For our numerical examples, we thus sweep through different values for the horizon H and report the lowest upper bound $R_H^* + (T - H)$.

VII. EXAMPLES

We demonstrate our approach on three examples: (i) the lava problem from the POMDP literature, (ii) an example with continuous state and observation spaces corresponding to a robot catching a freely-falling object, and (iii) obstacle avoidance using a depth sensor with non-Gaussian noise. We illustrate the strength of our upper bounds on these examples by comparing them against the performance achieved by concrete control policies (i.e., lower bounds on achievable performance). We also demonstrate the applicability of our approach for establishing the superiority of one sensor over another (from the perspective of a given task). Code for all examples can be found at: <https://github.com/irom-lab/performance-limits>.

A. Lava Problem

The first example we consider is the lava problem (Fig. 2) [26–28] from the POMDP literature.

Dynamics. The setting consists of five discrete states (Fig. 2) and two actions (left and right). If the robot falls into the lava state, it remains there (i.e., the lava state is absorbing). If the robot attempts to go left from state 1, it remains at state 1. The initial state distribution p_0 is chosen to be uniform over the non-lava states.

Sensor. The robot is equipped with a sensor that provides a (noisy) state estimate. The sensor reports the correct state (i.e.,

$o_t = s_t$) with probability p_{correct} (and a uniformly randomly chosen incorrect state with probability $1 - p_{\text{correct}}$).

Rewards. The robot's objective is to navigate to the goal state (which is an absorbing state), within a time horizon of $T = 5$. This objective is encoded via a reward function $r_t(s_t, a_t)$, which is purely state-dependent. The reward associated with being in the lava is 0; the reward associated with being at the goal is 1; the reward at all other states is 0.1.

Results. An interesting feature of this problem is that it admits a purely *open-loop* policy that achieves a high expected reward. In particular, consider the following sequence of actions: left, right, right. No matter which state the robot starts from, this sequence of actions will steer the robot to the goal state (recall that the goal is absorbing). Given the initial distribution and rewards above, this open-loop policy achieves an expected reward of 3.425. Suppose we set $p_{\text{correct}} = 1/5$ (i.e., the sensor just uniformly randomly returns a state estimate and thus provides no information regarding the state). In this case, Algorithm 1 returns an upper bound: $3.5 \geq R^*$.

Next, we plot the upper bounds provided by Algorithm 1 for different values of sensor noise by varying p_{correct} . The resulting bounds are shown in Fig. 3. Since the lava problem is a finite POMDP, one can compute R^* *exactly* using a POMDP solver. Fig. 3 compares the upper bounds on R^* returned by Algorithm 1 with R^* computed using the `pomdp_py` package [29]. The figure illustrates that our approach provides strong bounds on the best achievable reward for this problem. We also note that computation of the POMDP solution (for each value of p_{correct}) takes ~ 20 s, while the computation of the bound takes ~ 0.2 s (i.e., $\sim 100\times$ faster).

B. Catching a Falling Object

Next, we consider a problem with continuous state and observation spaces. The goal of the robot is to catch a freely-falling object such as a ball (Fig. 4).

Dynamics. We describe the four-dimensional state of the robot-ball system by $s_t := [x_t^{\text{rel}}, y_t^{\text{rel}}, v_t^{x,\text{ball}}, v_t^{y,\text{ball}}] \in \mathbb{R}^4$, where $[x_t^{\text{rel}}, y_t^{\text{rel}}]$ is the relative position of the ball with respect to the robot, and $[v_t^{x,\text{ball}}, v_t^{y,\text{ball}}]$ corresponds to the ball's velocity. The action a_t is the horizontal speed of the robot and can be chosen within the range $[-0.4, 0.4]$ m/s (discretized in increments of 0.1m/s). The dynamics of the system are given by:

$$s_{t+1} = \begin{bmatrix} x_{t+1}^{\text{rel}} \\ y_{t+1}^{\text{rel}} \\ v_{t+1}^{x,\text{ball}} \\ v_{t+1}^{y,\text{ball}} \end{bmatrix} = \begin{bmatrix} x_t^{\text{rel}} + (v_t^{x,\text{ball}} - a_t)\Delta t \\ y_t^{\text{rel}} + \Delta t v_t^{y,\text{ball}} \\ v_t^{x,\text{ball}} \\ v_t^{y,\text{ball}} - g\Delta t \end{bmatrix}, \quad (31)$$

where $\Delta t = 1$ is the time-step and $g = 0.1\text{m/s}^2$ is chosen such that the ball reaches the ground within a time horizon of $T = 5$. The initial state distribution p_0 is chosen to be a Gaussian with mean $[0.0\text{m}, 1.05\text{m}, 0.0\text{m/s}, 0.05\text{m/s}]$ and diagonal covariance matrix $\text{diag}([0.01^2, 0.1^2, 0.2^2, 0.1^2])$.

Sensor. The robot's sensor provides a noisy state estimate $o_t = s_t + \epsilon_t$, where ϵ_t is drawn from a Gaussian

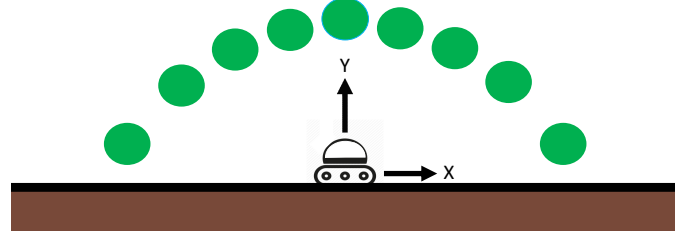


Fig. 4. An illustration of the ball-catching example with continuous state and observation spaces. The robot is constrained to move horizontally along the ground and can control its speed. Its goal is to track the position of the falling ball using a noisy estimate of the ball's state.

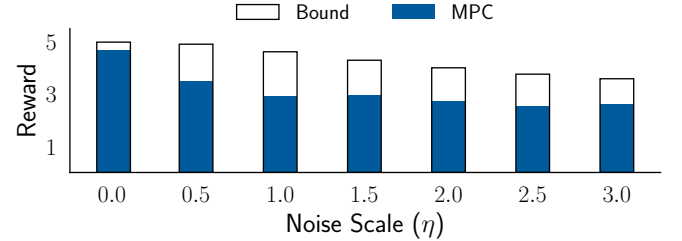


Fig. 5. Results for the ball-catching example. We compare the upper bounds on achievable expected rewards with the expected rewards using MPC combined with Kalman filtering for different values of sensor noise (results for MPC are averaged across five evaluation seeds; the std. dev. across seeds is too small to visualize).

distribution with zero mean and diagonal covariance matrix $\eta \cdot \text{diag}([0.5^2, 1.0^2, 0.75^2, 1.0^2])$. Here, η is a noise scale that we will vary in our experiments.

Rewards. The reward function $r_t(s_t, a_t)$ is chosen to encourage the robot to track the ball's motion. In particular, we choose $r_t(s_t, a_t) = \max(1 - 2|x_t^{\text{rel}}|, 0)$. The reward is large when x_{rel} is close to 0 (with a maximum reward of 1 when $x_t^{\text{rel}} = 0$); the robot receives no reward if $|x_t^{\text{rel}}| \geq 0.5$. The robot's goal is to maximize the expected cumulative reward over a time horizon of $T = 5$.

Results. Unlike the lava problem, this problem does not admit a good open-loop policy since the initial distribution on $v_0^{x,\text{ball}}$ is symmetric about 0; thus, the robot does not have *a priori* knowledge of the ball's x-velocity direction (as illustrated in Fig. 4). Fig. 5 plots the upper bound on the expected cumulative reward obtained using Algorithm 1 for different settings of the observation noise scale η . Since the dynamics (31) are affine, the sensor model is Gaussian, and the initial state distribution is also Gaussian, we can apply the techniques described in Sec. VI-A for *analytically* computing the quantities of interest in Algorithm 1.

Fig. 5 also compares the upper bounds on R^* with achievable *lower bounds* by applying a model-predictive control (MPC) scheme combined with a Kalman filter for state estimation. We estimate the expected reward achieved by the MPC controller using 100 initial conditions sampled from p_0 . Fig. 5 plots the average of these expected rewards across five random seeds (the resulting std. dev. is too small to visualize). As the figure illustrates, the MPC controller obeys the fundamental

bound on reward computed by our approach. Moreover, the performance of the controller qualitatively tracks the degradation of achievable performance predicted by the bound as η is increased. Finally, we observe that sensors with noise scales $\eta = 1$ and higher are *fundamentally limited* as compared to a noiseless sensor. This is demonstrated by the fact that the MPC controller for $\eta = 0$ achieves higher performance than the fundamental limit on performance for $\eta = 1$.

C. Obstacle Avoidance with a Depth Sensor

For our final example, we consider the problem of obstacle avoidance using a depth sensor (Fig. 1). This is a more challenging problem with higher-dimensional (continuous) state and observation spaces, and non-Gaussian sensor models.

State and action spaces. The robot is initialized at the origin with six cylindrical obstacles of fixed radius placed randomly in front of it. The state $s_t \in \mathbb{R}^{12}$ of this system describes the locations of these obstacles in the environment. In addition, we also place “walls” enclosing a workspace $[-1, 1]\text{m} \times [-0.1, 1.2]\text{m}$ (these are not visualized in the figure to avoid clutter). The initial state distribution p_0 corresponds to uniformly randomly choosing the x-y locations of the cylindrical obstacles from the set $[-1, 1]\text{m} \times [0.9, 1.1]\text{m}$. The robot’s goal is to navigate to the end of the workspace by choosing a motion primitive to execute (based on a noisy depth sensor described below). Fig. 1 illustrates the set of ten motion primitives the robot can choose from; this set corresponds to the action space.

Rewards. We treat this problem as a one-step decision making problem (Sec. IV). Once the robot chooses a motion primitive to execute based on its sensor measurements, it receives a reward of 0 if the motion primitive results in a collision with an obstacle; if the motion primitive results in collision-free motion, the robot receives a reward of 1. The expected reward for this problem thus corresponds to the probability of safe (i.e., collision-free) motion.

Sensor. The robot is equipped with a depth sensor which provides distances along $n_{\text{rays}} = 10$ rays. The sensor has a field of view of 90° and a maximum range of 1.5m. We use the noise model for range finders described in [22, Ch. 6.3] and consider two kinds of measurement errors: (i) errors due to failures to detect obstacles, and (ii) random noise in the reported distances. For each ray, there is a probability $p_{\text{miss}} = 0.1$ that the sensor misses an obstacle and reports the maximum range (1.5m) instead. In the case that an obstacle is not missed, the distance reported along a given ray is sampled from a Gaussian with mean equal to the true distance along that ray and std. dev. equal to η . The noise for each ray is sampled independently. Overall, this is a non-Gaussian sensor model due to the combination of the two kinds of errors.

Results. We implement Algorithm 1 using the sampling-based techniques described in Sec. VI-B. We sample 20K obstacle environments for upper bounding the open-loop rewards associated with each action. We also utilize 20K batches (each of size $K = 1000$) for upper bounding the mutual information using (29). We utilize a version of Hoeffding’s

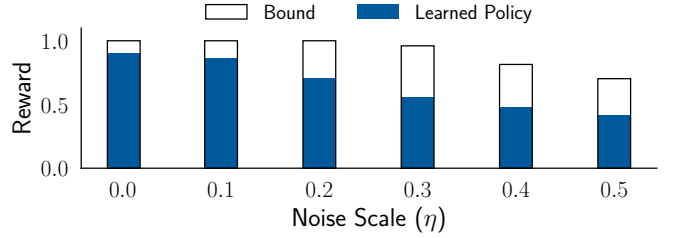


Fig. 6. Results for the obstacle avoidance example. We compare the upper bounds on achievable expected rewards with the expected rewards using a learned neural network policy for different values of sensor noise (results for the learned policy are averaged across five training seeds; the std. dev. is too small to visualize).

inequality (Theorem 3) presented in Appendix B to obtain an upper bound on R^* that holds with probability $1 - \delta = 0.95$.

Fig. 6 shows the resulting upper bounds for different values of the observation noise std. dev. η . We compare these with rewards achieved by neural network policies trained to perform this task. For each η , we sample 5000 training environments and corresponding sensor observations. For each environment, we generate a ten-dimensional training label by recording the minimum distance to the obstacles achieved by each motion primitive and passing the vector of distances through an (element-wise) softmax transformation. We use a cross-entropy loss to train a neural network that predicts the label of distances for each primitive given a sensor observation as input. We use two fully-connected layers with a ReLU nonlinearity; the output is passed through a softmax layer. At test-time, a given sensor observation is passed as input to the trained neural network; the motion primitive corresponding to the highest predicted distance is then executed. We estimate the expected reward achieved by the trained policy using 5000 test environments (unseen during training). Fig. 6 plots the average of these expected rewards across five training seeds for each value of η (the std. dev. across seeds is too small to visualize).

We emphasize that the upper bound on R^* computed using our approach is a fundamental limit on performance that holds regardless of the size of the neural network, the network architecture, or algorithm used for policy learning. We also highlight the fact that a sensor with noise $\eta = 0.1$ achieves higher performance than the fundamental limit for a sensor with noise $\eta = 0.4$ or 0.5 .

Finally, we apply our approach in order to compare two sensors with varying number n_{rays} of rays along which the depth sensor provides distance estimates (Fig. 1). For this comparison, we fix $\eta = 0.3$ and $p_{\text{miss}} = 0.05$. We compare two sensors with $n_{\text{rays}} = 50$ (Sensor 1) and $n_{\text{rays}} = 5$ (Sensor 2) respectively. The upper bound on expected reward computed using our approach (with confidence $1 - \delta = 0.95$) for Sensor 2 is 0.79. A neural network policy for Sensor 1 achieves an expected reward of approximately 0.86, which surpasses the fundamental limit on performance for Sensor 2.

VIII. DISCUSSION AND CONCLUSIONS

We have presented an approach for establishing fundamental limits on performance for sensor-based robot control and policy

learning. We defined a quantity that captures the amount of task-relevant information provided by a sensor; using a novel version of the generalized Fano inequality, we demonstrated that this quantity upper bounds the expected reward for one-step decision making problems. We developed a dynamic programming approach for extending this bound to multi-step problems. The resulting framework has potentially broad applicability to robotic systems with continuous state and observation spaces, nonlinear and stochastic dynamics, and non-Gaussian sensor models. Our numerical experiments demonstrate the ability of our approach to establish strong bounds on performance for such settings. In addition, we provided an application of our approach for comparing different sensors and establishing the superiority of one sensor over another for a given task.

Challenges and future work. There are a number of challenges and directions for future work associated with our approach. On the theoretical front, we expect that our bounds in Theorems 1 and 2 can be extended to utilize general f-divergences (instead of the KL divergence); this could potentially lead to tighter bounds. In addition, it would be interesting to handle settings where the sensor model is inaccurate (in contrast to this paper, where we have focused on establishing fundamental limits given a particular sensor model). For example, one could potentially perform an adversarial search over a family of sensor models in order to find the model that results in the lowest bound on achievable performance.

On the algorithmic front, the primary challenges are: (i) efficient computation of bounds for longer time horizons, and (ii) extensions to continuous action spaces. As presented, Algorithm 1 requires an enumeration over action sequences. Finding more computationally efficient versions of Algorithm 1 is thus an important direction for future work. The primary bottleneck in extending our approach to continuous action spaces is the need to perform a supremization over actions when computing $\bar{R}_t^{\perp*}$ in Algorithm 1. However, we note that any upper bound on $\bar{R}_t^{\perp*}$ also leads to a valid upper bound on R^* . Thus, one possibility is to use a Lagrangian relaxation [18] to upper bound $\bar{R}_t^{\perp*}$ in settings with continuous action spaces.

Our work also opens up a number of exciting directions for longer-term research. While we have focused on establishing fundamental limits imposed by imperfect sensing in this work, one could envision a broader research agenda that seeks to establish bounds on performance due to other limited resources (e.g., onboard computation or memory). One concrete direction is to combine the techniques presented here with information-theoretic notions of bounded rationality [30–32]. Finally, another exciting direction is to turn the impossibility results provided by our approach into certificates of *robustness* against an adversary. Specifically, consider an adversary that can observe our robot’s actions; if one could establish fundamental limits on performance *for the adversary* due to its inability to infer the robot’s internal state (and hence its future behavior) using past observations, this provides a certificate of robustness against *any* adversary. This is reminiscent of a long tradition in cryptography of turning impossibility or hardness results

into robust protocols for security [33, Ch. 9].

Overall, we believe that the ability to establish fundamental limits on performance for robotic systems is crucial for establishing a science of robotics. We hope that the work presented here along with the indicated directions for future work represent a step in this direction.

ACKNOWLEDGMENTS

The authors were partially supported by the NSF CAREER Award [#2044149] and the Office of Naval Research [N00014-21-1-2803]. The authors would like to thank Alec Farid and David Snyder for helpful discussions on this work.

REFERENCES

- [1] Oliver Brock. Is robotics in need of a paradigm shift? In *Berlin Summit on Robotics: Conference Report*, pages 1–10, 2011.
- [2] James Morris. Can Tesla really do without radar for full self-driving? *Forbes*, Jul 2021.
- [3] Daniel E. Koditschek. What is robotics? Why do we need it and how can we get it? *Annual Review of Control, Robotics, and Autonomous Systems*, 4:1–33, 2021.
- [4] Sertac Karaman and Emilio Frazzoli. High-speed flight in an ergodic forest. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 2899–2906, 2012.
- [5] Sanjiban Choudhury, Sebastian Scherer, and J. Andrew Bagnell. Theoretical limits of speed and resolution for kinodynamic planning in a Poisson forest. In *Proceedings of Robotics: Science and Systems (RSS)*, 2015.
- [6] Béla Bollobás and Oliver Riordan. *Percolation*. Cambridge University Press, 2006.
- [7] Davide Falanga, Suseong Kim, and Davide Scaramuzza. How fast is too fast? The role of perception latency in high-speed sense and avoid. *IEEE Robotics and Automation Letters*, 4(2):1884–1891, 2019.
- [8] John C. Doyle, Bruce A. Francis, and Allen R. Tannenbaum. *Feedback Control Theory*. Courier Corporation, 2013.
- [9] Yoke Peng Leong and John C. Doyle. Understanding robust control theory via stick balancing. In *Proceedings of the IEEE Conference on Decision and Control (CDC)*, pages 1508–1514, 2016.
- [10] Jingxi Xu, Bruce Lee, Nikolai Matni, and Dinesh Jayaraman. How are learned perception-based controllers impacted by the limits of robust control? In *Learning for Dynamics and Control*, pages 954–966. PMLR, 2021.
- [11] Steven M. LaValle. *Sensing and filtering: A fresh perspective based on preimages and information spaces*. Citeseer, 2012.
- [12] Steven M. LaValle. Sensor lattices: Structures for comparing information feedback. In *International Workshop on Robot Motion and Control (RoMoCo)*, pages 239–246, 2019.

- [13] Jason M. O’Kane and Steven M. LaValle. Comparing the power of robots. *The International Journal of Robotics Research*, 27(1):5–23, 2008.
- [14] Thomas M. Cover. *Elements of Information Theory*. John Wiley & Sons, 1999.
- [15] John C. Duchi and Martin J. Wainwright. Distance-based and continuum Fano inequalities with applications to statistical estimation. *arXiv preprint arXiv:1311.2669*, 2013.
- [16] Xi Chen, Adityanand Guntuboyina, and Yuchen Zhang. On Bayes risk lower bounds. *The Journal of Machine Learning Research*, 17(1):7687–7744, 2016.
- [17] Sebastien Gerchinovitz, Pierre Ménard, and Gilles Stoltz. Fano’s inequality for random variables. *Statistical Science*, 35(2):178–201, 2020.
- [18] Stephen P. Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- [19] Paul Milgrom and Ilya Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2):583–601, 2002.
- [20] MOSEK ApS. MOSEK Fusion API for Python 9.0.84 (beta), 2019. URL <https://docs.mosek.com/9.0/pythonfusion/index.html>.
- [21] B. O’Donoghue, E. Chu, N. Parikh, and S. Boyd. SCS: Splitting conic solver, version 2.0.2. <https://github.com/cvxgrp/scs>, November 2017.
- [22] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic Robotics*. MIT Press, 2005.
- [23] John C. Duchi. Derivations for linear algebra and optimization. 2016.
- [24] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.
- [25] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pages 5171–5180. PMLR, 2019.
- [26] Anthony R. Cassandra, Leslie P. Kaelbling, and Michael L. Littman. Acting optimally in partially observable stochastic domains. In *AAAI*, volume 94, pages 1023–1028, 1994.
- [27] Peter R. Florence. Integrated perception and control at high speed. Master’s thesis, Massachusetts Institute of Technology, 2017.
- [28] Vincent Pacelli and Anirudha Majumdar. Learning task-driven control policies via information bottlenecks. In *Proceedings of Robotics: Science and Systems (RSS)*, 2020.
- [29] Kaiyu Zheng and Stefanie Tellex. Pomdp_py: A framework to build and solve POMDP problems. *arXiv preprint arXiv:2004.10099*, 2020.
- [30] Naftali Tishby and Daniel Polani. Information theory of decisions and actions. In *Perception-Action Cycle: Models, Architectures, and Hardware*, pages 601–636. Springer, 2011.
- [31] Pedro A. Ortega, Daniel A. Braun, Justin Dyer, Kee-Eung Kim, and Naftali Tishby. Information-theoretic bounded rationality. *arXiv preprint arXiv:1512.06789*, 2015.
- [32] Vincent Pacelli and Anirudha Majumdar. Robust control under uncertainty via bounded rationality and differential privacy. *arXiv preprint arXiv:2109.08262*, 2021.
- [33] Jean-Philippe Aumasson. *Serious Cryptography: A Practical Introduction to Modern Encryption*. No Starch Press, 2017.
- [34] Wolfgang Mulzer. Five proofs of Chernoff’s bound with applications. *arXiv preprint arXiv:1801.03365*, 2018.

APPENDIX A
PROOFS

Lemma 4 (Monotonicity of KL inverse). $\mathbb{D}^{-1}(q|c)$ is monotonically non-decreasing in $q \in [0, 1]$ for $c \geq 0$.

Proof: Recall that the KL inverse is defined using the following optimization problem with decision variable p :

$$\mathbb{D}^{-1}(q|c) := \sup \{p \in [0, 1] \mid \mathbb{D}_{\mathcal{B}}(p||q) \leq c\}. \quad (32)$$

We first prove that:

$$\mathbb{D}^{-1}(q|c) \geq q. \quad (33)$$

For contradiction, suppose that $\mathbb{D}^{-1}(q|c) < q$. Then, let $p' := q$. Notice that p' is a feasible point for the optimization problem (32) since $\mathbb{D}_{\mathcal{B}}(p'||q) = \mathbb{D}_{\mathcal{B}}(q||q) = 0$, and $p' = q \in [0, 1]$. But, by assumption, $p' = q > \mathbb{D}^{-1}(q|c)$; this contradicts the statement that $\mathbb{D}^{-1}(q|c)$ is the optimal solution to (32).

This allows us to demonstrate the monotonicity of $\mathbb{D}^{-1}(q|c)$ in q for the case when $c = 0$. In particular, when $c = 0$, we have that $\mathbb{D}^{-1}(q|c) = q$. Thus, monotonicity of $\mathbb{D}^{-1}(q|c)$ in q follows immediately.

Next, we demonstrate the monotonicity of $\mathbb{D}^{-1}(q|c)$ in q for the case when $c > 0$. We note that:

- The objective function of (32) is continuous and concave (indeed, linear).
- The constraints of (32) can be written in the form $g(p, q) \geq 0$ with g continuous and concave (since the KL divergence is jointly convex in both arguments, and the other constraints on p are linear).
- There is a point (in particular, the point $p = q$) that is strictly feasible (i.e., $g(p, q) > 0$) for (32) for $c > 0$ and $q \in [0, 1]$.

The optimization problem (32) thus satisfies the conditions for the envelope theorem [19, Corollary 5], which allows us to characterize the derivative of (32) with respect to the parameter q :

$$\frac{\partial[\mathbb{D}^{-1}(q|c)]}{\partial q} = \sup_{p \in X^*(q)} \inf_{\lambda \in Y^*(q)} \frac{\partial L(p, q, \lambda)}{\partial q}, \quad (34)$$

where $L(p, q, \lambda)$ is the Lagrangian for (32), $X^*(q)$ are the set of optimizers for (32), $Y^*(q)$ are the set of optimizers for the Lagrange dual, and $\lambda = [\lambda_1, \lambda_2, \lambda_3]$ are the Lagrange multipliers corresponding to the three constraints:

$$c - \mathbb{D}_{\mathcal{B}}(p||q) \geq 0, \quad (35)$$

$$p \geq 0, \quad (36)$$

$$1 - p \geq 0. \quad (37)$$

The Lagrangian is given by:

$$L(p, q, \lambda) = p + \lambda_1(c - \mathbb{D}_{\mathcal{B}}(p||q)) + \lambda_2 p + \lambda_3(1 - p). \quad (38)$$

Computing the partial derivative with respect to q :

$$\frac{\partial L(p, q, \lambda)}{\partial q} = -\lambda_1 \frac{\partial \mathbb{D}_{\mathcal{B}}(p||q)}{\partial q} \quad (39)$$

$$= -\lambda_1 \frac{\partial}{\partial q} \left[p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q} \right] \quad (40)$$

$$= -\lambda_1 \left[-\frac{p}{q} + \frac{1 - p}{1 - q} \right]. \quad (41)$$

We note that the Lagrange multiplier λ_1 must be nonnegative. In addition, we know from (33) that the optimal value p^* of the optimization problem (32) must be greater than or equal to q . Thus:

$$\frac{p^*}{q} \geq 1, \text{ and } \frac{1 - p^*}{1 - q} \leq 1 \quad (42)$$

$$\implies \frac{p^*}{q} \geq \frac{1 - p^*}{1 - q} \quad (43)$$

$$\implies \frac{\partial L(p, q, \lambda)}{\partial q} \geq 0. \quad (44)$$

Hence, the envelope theorem (34) demonstrates that $\mathbb{D}^{-1}(q|c)$ is monotonically non-decreasing in q . ■

Theorem 1 (Single-step performance bound). *The best achievable reward is upper bounded by the task-relevant information potential (TRIP) of a sensor:*

$$\tau(\sigma_0; r_0) \geq R^*(\sigma_0; r_0) := \sup_{\pi_0} \mathbb{E}_{s_0, o_0} [r_0(s_0, \pi_0(o_0))]. \quad (9)$$

Proof: For a given policy π_0 , define:

$$R := \mathbb{E}_{p_0(s_0)} \mathbb{E}_{\sigma_0(o_0|s_0)} [r_0(s_0, \pi_0(o_0))], \quad (45)$$

and

$$\tilde{R} := \mathbb{E}_{p_0(s_0)} \mathbb{E}_{q(o_0)} [r_0(s_0, \pi_0(o_0))]. \quad (46)$$

The only difference between R and $\tilde{R}$ is that the observations o_0 in $\tilde{R}$ are drawn from a state-independent distribution q .

Now, assuming bounded rewards (Assumption 1), we have:

$$\mathbb{D}_{\mathcal{B}}(R \| \tilde{R}) = \mathbb{D}_{\mathcal{B}} \left(\mathbb{E}_{p_0(s_0)} \mathbb{E}_{\sigma_0(o_0|s_0)} [r_0(s_0, \pi_0(o_0))] \parallel \mathbb{E}_{p_0(s_0)} \mathbb{E}_{q(o_0)} [r_0(s_0, \pi_0(o_0))] \right) \quad (47)$$

$$\leq \mathbb{E}_{p_0(s_0)} \mathbb{D}_{\mathcal{B}} \left(\mathbb{E}_{\sigma_0(o_0|s_0)} [r_0(s_0, \pi_0(o_0))] \parallel \mathbb{E}_{q(o_0)} [r_0(s_0, \pi_0(o_0))] \right) \quad (48)$$

$$\leq \mathbb{E}_{p_0(s_0)} \mathbb{D}(\sigma_0(o_0|s_0) \parallel q(o_0)). \quad (49)$$

The first inequality above follows from Jensen's inequality, while the second follows from the data processing inequality (see [17, Corollary 2] for the specific version). From the well-known variational representation of the mutual information (see, e.g., [16]), we note that:

$$\mathbb{I}(o_0; s_0) = \inf_q \mathbb{E}_{p_0(s_0)} \mathbb{D}(\sigma_0(o_0|s_0) \parallel q(o_0)). \quad (50)$$

Thus, taking the infimum of (49) with respect to q , we see:

$$\mathbb{D}_{\mathcal{B}}(R \| \tilde{R}) \leq \mathbb{I}(o_0; s_0). \quad (51)$$

We can then upper bound R using the KL inverse:

$$R \leq \mathbb{D}^{-1}(\tilde{R} \mid \mathbb{I}(o_0; s_0)). \quad (52)$$

Thus,

$$\sup_{\pi_0} R \leq \sup_{\pi_0} \mathbb{D}^{-1}(\tilde{R} \mid \mathbb{I}(o_0; s_0)). \quad (53)$$

From the monotonicity of the KL inverse (Lemma 4), we have:

$$\sup_{\pi_0} R \leq \mathbb{D}^{-1} \left(\sup_{\pi_0} \tilde{R} \mid \mathbb{I}(o_0; s_0) \right). \quad (54)$$

Since by definition,

$$R^*(\sigma_0; r_0) = \sup_{\pi_0} R, \quad (55)$$

we have:

$$R^*(\sigma_0; r_0) \leq \mathbb{D}^{-1} \left(\sup_{\pi_0} \tilde{R} \mid \mathbb{I}(o_0; s_0) \right). \quad (56)$$

Now, using the Fubini-Tonelli theorem, we see:

$$\sup_{\pi_0} \tilde{R} = \sup_{\pi_0} \mathbb{E}_{p_0(s_0)} \mathbb{E}_{q(o_0)} [r_0(s_0, \pi_0(o_0))] \quad (57)$$

$$= \sup_{\pi_0} \mathbb{E}_{q(o_0)} \mathbb{E}_{p_0(s_0)} [r_0(s_0, \pi_0(o_0))] \quad (58)$$

$$\leq \mathbb{E}_{q(o_0)} \sup_{\pi_0} \mathbb{E}_{p_0(s_0)} [r_0(s_0, \pi_0(o_0))] \quad (59)$$

$$= \mathbb{E}_{q(o_0)} \sup_{a_0} \mathbb{E}_{p_0(s_0)} [r_0(s_0, a_0)] \quad (60)$$

$$= \sup_{a_0} \mathbb{E}_{p_0(s_0)} [r_0(s_0, a_0)]. \quad (61)$$

Since open-loop actions are special cases of policies, we also have:

$$\sup_{\pi_0} \tilde{R} = \sup_{\pi_0} \mathbb{E}_{p_0(s_0)} \mathbb{E}_{q(o_0)} [r_0(s_0, \pi_0(o_0))] \geq \sup_{a_0} \mathbb{E}_{p_0(s_0)} [r_0(s_0, a_0)]. \quad (62)$$

As a result, we see that:

$$\sup_{\pi_0} \tilde{R} = \sup_{a_0} \mathbb{E}_{p_0(s_0)} [r_0(s_0, a_0)] = R_0^\perp, \quad (63)$$

where $R_0^\perp$ is as defined in (7). Combining this with (56), we obtain the desired result:

$$R^*(\sigma_0; r_0) \leq \mathbb{D}^{-1} \left(R_0^\perp \mid \mathbb{I}(o_0; s_0) \right) =: \tau(\sigma_0; r_0). \quad (64)$$

■

Proposition 2 (Recursive bound). *For any $t = 0, \dots, T-1$, the following inequality holds for any open-loop sequence of actions $a_{0:t-1}$:*

$$\sup_{\pi_t^t, \dots, \pi_{T-1}^t} R_t \leq \underbrace{(T-t) \cdot \mathbb{D}^{-1} \left(\frac{R_t^{\perp*}}{T-t} \mid \mathbb{I}(o_t; s_t) \right)}_{=: \tau_t(\sigma_{t:T-1}; r_{t:T-1})}. \quad (15)$$

Proof: The proof follows a similar structure to that of Theorem 1. First, note that R_t defined in (10) can be written as:

$$R_t = \mathbb{E}_{s_t | a_{0:t-1}} \mathbb{E}_{o_t | s_t} \mathbb{E}_{s_{t+1:T-1}, o_{t+1:T-1} | s_t, o_t} \left[\sum_{k=t}^{T-1} r_k(s_k, \pi_k^t(o_{t:k})) \right].$$

Define:

$$\tilde{R}_t := \mathbb{E}_{s_t | a_{0:t-1}} \mathbb{E}_{q(o_t)} \mathbb{E}_{s_{t+1:T-1}, o_{t+1:T-1} | s_t, o_t} \left[\sum_{k=t}^{T-1} r_k(s_k, \pi_k^t(o_{t:k})) \right].$$

The only difference between $\tilde{R}_t$ and R_t is that the observations o_t in $\tilde{R}_t$ are drawn from a state-independent distribution q .

For the sake of notational simplicity, we will assume that R_t and $\tilde{R}_t$ have been normalized to be within $[0, 1]$ by scaling with $1/(T-t)$. The desired result (15) then follows from the bound we prove below by simply rescaling with $(T-t)$.

Now,

$$\begin{aligned} \mathbb{D}_{\mathcal{B}}(R_t \parallel \tilde{R}_t) &= \mathbb{D}_{\mathcal{B}} \left(\mathbb{E}_{s_t | a_{0:t-1}} \mathbb{E}_{o_t | s_t} \mathbb{E}_{s_{t+1:T-1}, o_{t+1:T-1} | s_t, o_t} \left[\sum_{k=t}^{T-1} r_k(s_k, \pi_k^t(o_{t:k})) \right] \parallel \mathbb{E}_{s_t | a_{0:t-1}} \mathbb{E}_{q(o_t)} \mathbb{E}_{s_{t+1:T-1}, o_{t+1:T-1} | s_t, o_t} \left[\sum_{k=t}^{T-1} r_k(s_k, \pi_k^t(o_{t:k})) \right] \right) \\ &\leq \mathbb{E}_{s_t | a_{0:t-1}} \mathbb{D}_{\mathcal{B}} \left(\mathbb{E}_{o_t | s_t} \mathbb{E}_{s_{t+1:T-1}, o_{t+1:T-1} | s_t, o_t} \left[\sum_{k=t}^{T-1} r_k(s_k, \pi_k^t(o_{t:k})) \right] \parallel \mathbb{E}_{q(o_t)} \mathbb{E}_{s_{t+1:T-1}, o_{t+1:T-1} | s_t, o_t} \left[\sum_{k=t}^{T-1} r_k(s_k, \pi_k^t(o_{t:k})) \right] \right) \\ &\leq \mathbb{E}_{s_t | a_{0:t-1}} \mathbb{D} \left(\sigma_t(o_t | s_t) \parallel q(o_t) \right). \end{aligned} \quad (65)$$

The first inequality above follows from Jensen's inequality, while the second follows from the data processing inequality (see [17, Corollary 2] for the specific version).

From the well-known variational representation of the mutual information [16], we note that:

$$\mathbb{I}(o_t; s_t) = \inf_q \mathbb{E}_{s_t | a_{0:t-1}} \mathbb{D} \left(\sigma_t(o_t | s_t) \parallel q(o_t) \right). \quad (66)$$

Thus, taking the infimum of (65) with respect to q , we see:

$$\mathbb{D}_{\mathcal{B}}(R_t \parallel \tilde{R}_t) \leq \mathbb{I}(o_t; s_t). \quad (67)$$

We can then upper bound R_t using the KL inverse:

$$R_t \leq \mathbb{D}^{-1} \left(\tilde{R}_t \mid \mathbb{I}(o_t; s_t) \right). \quad (68)$$

Thus,

$$\sup_{\pi_t^t, \dots, \pi_{T-1}^t} R_t \leq \sup_{\pi_t^t, \dots, \pi_{T-1}^t} \mathbb{D}^{-1} \left(\tilde{R}_t \mid \mathbb{I}(o_t; s_t) \right). \quad (69)$$

Notice that the LHS is precisely the quantity we are interested in upper bounding in Proposition 2. From the monotonicity of the KL inverse (Lemma 4), we have:

$$\sup_{\pi_t^t, \dots, \pi_{T-1}^t} R_t \leq \mathbb{D}^{-1}\left(\tilde{R}_t^\star \mid \mathbb{I}(o_t; s_t)\right), \quad (70)$$

where

$$\tilde{R}_t^\star := \sup_{\pi_t^t, \dots, \pi_{T-1}^t} \tilde{R}_t = \sup_{\pi_t^t, \dots, \pi_{T-1}^t} \mathbb{E}_{a_{0:t-1}} \mathbb{E}_{s_t | q(o_t)} \mathbb{E}_{s_{t+1:T-1}, o_{t+1:T-1} | s_t, o_t} \left[\sum_{k=t}^{T-1} r_k(s_k, \pi_k^t(o_{t:k})) \right]. \quad (71)$$

Now, using the Fubini-Tonelli theorem:

$$\sup_{\pi_t^t, \dots, \pi_{T-1}^t} \mathbb{E}_{a_{0:t-1}} \mathbb{E}_{s_t | q(o_t)} \mathbb{E}_{s_{t+1:T-1}, o_{t+1:T-1} | s_t, o_t} \left[\sum_{k=t}^{T-1} r_k(s_k, \pi_k^t(o_{t:k})) \right] \quad (72)$$

$$= \sup_{\pi_t^t, \dots, \pi_{T-1}^t} \mathbb{E}_{q(o_t)} \mathbb{E}_{s_t | a_{0:t-1}} \mathbb{E}_{s_{t+1:T-1}, o_{t+1:T-1} | s_t, o_t} \left[\sum_{k=t}^{T-1} r_k(s_k, \pi_k^t(o_{t:k})) \right] \quad (73)$$

$$= \sup_{\pi_{t+1}^t, \dots, \pi_{T-1}^t} \left[\sup_{\pi_t^t} \mathbb{E}_{q(o_t)} \mathbb{E}_{s_t | a_{0:t-1}} \mathbb{E}_{s_{t+1:T-1}, o_{t+1:T-1} | s_t, o_t} \left[\sum_{k=t}^{T-1} r_k(s_k, \pi_k^t(o_{t:k})) \right] \right] \quad (74)$$

$$\leq \sup_{\pi_{t+1}^t, \dots, \pi_{T-1}^t} \left[\mathbb{E}_{q(o_t)} \sup_{\pi_t^t} \mathbb{E}_{s_t | a_{0:t-1}} \mathbb{E}_{s_{t+1:T-1}, o_{t+1:T-1} | s_t, o_t} \left[\sum_{k=t}^{T-1} r_k(s_k, \pi_k^t(o_{t:k})) \right] \right]. \quad (75)$$

Notice that:

$$\mathbb{E}_{q(o_t)} \sup_{\pi_t^t} \mathbb{E}_{s_t | a_{0:t-1}} \mathbb{E}_{s_{t+1:T-1}, o_{t+1:T-1} | s_t, o_t} \left[\sum_{k=t}^{T-1} r_k(s_k, \pi_k^t(o_{t:k})) \right] \quad (76)$$

$$= \mathbb{E}_{q(o_t)} \sup_{\pi_t^t} \mathbb{E}_{s_t | a_{0:t-1}} \left[r_t(s_t, \pi_t^t(o_t)) + \mathbb{E}_{s_{t+1}, o_{t+1} | s_t, \pi_t^t(o_t)} \left[r_{t+1}(s_{t+1}, \pi_{t+1}^t(o_{t:t+1})) + \dots \right] \dots \right] \quad (77)$$

$$= \mathbb{E}_{q(o_t)} \sup_{a_t} \mathbb{E}_{s_t | a_{0:t-1}} \left[r_t(s_t, a_t) + \underbrace{\mathbb{E}_{s_{t+1}, o_{t+1} | s_t, a_t} \left[r_{t+1}(s_{t+1}, \pi_{t+1}^{t+1}(o_{t+1})) + \dots \right] \dots}_{\text{Does not depend on } o_t} \right] \quad (78)$$

$$= \sup_{a_t} \mathbb{E}_{s_t | a_{0:t-1}} \left[r_t(s_t, a_t) + \mathbb{E}_{s_{t+1}, o_{t+1} | s_t, a_t} \left[r_{t+1}(s_{t+1}, \pi_{t+1}^{t+1}(o_{t+1})) + \dots \right] \dots \right]. \quad (79)$$

Here, (78) follows (77) since q is a fixed distribution that does not depend on the state.

We thus see that (75) equals:

$$\sup_{\pi_{t+1}^t, \dots, \pi_{T-1}^t} \left[\sup_{a_t} \mathbb{E}_{s_t | a_{0:t-1}} \left[r_t(s_t, a_t) + \underbrace{\mathbb{E}_{s_{t+1}, o_{t+1} | s_t, a_t} \left[r_{t+1}(s_{t+1}, \pi_{t+1}^{t+1}(o_{t+1})) + \dots \right] \dots}_{\text{Does not depend on } o_t} \right] \right] \quad (80)$$

$$= \sup_{\pi_{t+1}^{t+1}, \dots, \pi_{T-1}^{t+1}} \left[\sup_{a_t} \mathbb{E}_{s_t | a_{0:t-1}} \left[r_t(s_t, a_t) + \mathbb{E}_{s_{t+1}, o_{t+1} | s_t, a_t} \left[r_{t+1}(s_{t+1}, \pi_{t+1}^{t+1}(o_{t+1})) + \dots \right] \dots \right] \right] \quad (81)$$

$$= \sup_{\pi_{t+1}^{t+1}, \dots, \pi_{T-1}^{t+1}} \left[\sup_{a_t} \mathbb{E}_{s_t | a_{0:t-1}} \left[r_t(s_t, a_t) \right] + \mathbb{E}_{s_{t+1:T-1}, o_{t+1:T-1} | a_{0:t}} \left[\sum_{k=t+1}^{T-1} r_k(s_k, \pi_k^{t+1}(o_{t+1:k})) \right] \right] \quad (82)$$

$$= \sup_{\pi_{t+1}^{t+1}, \dots, \pi_{T-1}^{t+1}} R_t^\perp \quad (83)$$

$$= R_t^{\perp \star}. \quad (84)$$

We have thus proved that $\tilde{R}_t^* \leq R_t^{\perp*}$ (indeed, since open-loop policies are special cases of feedback policies, we also have $\tilde{R}_t^* \geq R_t^{\perp*}$ and thus $\tilde{R}_t^* = R_t^{\perp*}$). Since the RHS of (70) is a monotonically non-decreasing function of $\tilde{R}_t^*$ (Lemma 4), the observation that $\tilde{R}_t^* \leq R_t^{\perp*}$ implies:

$$\sup_{\pi_t^t, \dots, \pi_{T-1}^t} R_t \leq \mathbb{D}^{-1}\left(R_t^{\perp*} \mid \mathbb{I}(o_t; s_t)\right). \quad (85)$$

■

Theorem 2 (Multi-step performance bound). *Algorithm 1 returns an upper bound on the best achievable reward R^* .*

Proof: Using (backwards) induction, we prove that for all $t = T-1, \dots, 0$,

$$\sup_{\pi_t^t, \dots, \pi_{T-1}^t} R_t \leq \bar{R}_t(a_{0:t-1}), \quad \forall a_{0:t-1}. \quad (86)$$

Thus, in particular,

$$R^* = \sup_{\pi_0^0, \dots, \pi_{T-1}^0} R_0 \leq \bar{R}_0. \quad (87)$$

We prove (86) by backwards induction starting from $t = T-1$. In particular, Proposition 2 leads to the inductive step. We first prove the base step of induction using $t = T-1$. Using (15), we see:

$$\sup_{\pi_{T-1}^{T-1}} R_{T-1} \leq \mathbb{D}^{-1}\left(R_{T-1}^{\perp*} \mid \mathbb{I}(o_{T-1}; s_{T-1})\right). \quad (88)$$

By definition (see (14)),

$$\begin{aligned} R_{T-1}^{\perp*} &= \sup_{a_{T-1}} \mathbb{E}_{s_{T-1}|a_{0:T-2}} \left[r_{T-1}(s_{T-1}, a_{T-1}) \right] + \underbrace{R_T}_{=0} \\ &= \sup_{a_{T-1}} \mathbb{E}_{s_{T-1}|a_{0:T-2}} \left[r_{T-1}(s_{T-1}, a_{T-1}) \right] \\ &= \bar{R}_{T-1}^{\perp*}. \end{aligned}$$

Combining this with (88) and the monotonicity of the KL inverse (Proposition 1), we see:

$$\sup_{\pi_{T-1}^{T-1}} R_{T-1} \leq \mathbb{D}^{-1}\left(\bar{R}_{T-1}^{\perp*} \mid \mathbb{I}(o_{T-1}; s_{T-1})\right) \quad (89)$$

$$= \bar{R}_{T-1}(a_{0:T-2}). \quad (90)$$

Next, we prove the induction step. Suppose it is the case that for $t \in \{0, \dots, T-2\}$, we have

$$\sup_{\pi_{t+1}^{t+1}, \dots, \pi_{T-1}^{t+1}} R_{t+1} \leq \bar{R}_{t+1}(a_{0:t}). \quad (91)$$

We then need to show that

$$\sup_{\pi_t^t, \dots, \pi_{T-1}^t} R_t \leq \bar{R}_t(a_{0:t-1}). \quad (92)$$

To prove this, we first observe that

$$\begin{aligned} R_t^{\perp*} &:= \sup_{\pi_{t+1}^{t+1}, \dots, \pi_{T-1}^{t+1}} \sup_{a_t} \left[\mathbb{E}_{s_t|a_{0:t-1}} \left[r_t(s_t, a_t) \right] + R_{t+1} \right] \\ &= \sup_{a_t} \left[\mathbb{E}_{s_t|a_{0:t-1}} \left[r_t(s_t, a_t) \right] + \sup_{\pi_{t+1}^{t+1}, \dots, \pi_{T-1}^{t+1}} R_{t+1} \right]. \end{aligned}$$

Combining this with the induction hypothesis (91), we see

$$R_t^{\perp*} \leq \sup_{a_t} \left[\mathbb{E}_{s_t|a_{0:t-1}} \left[r_t(s_t, a_t) \right] + \bar{R}_{t+1}(a_{0:t}) \right] \quad (93)$$

$$=: \bar{R}_t^{\perp*}. \quad (94)$$

Finally, combining this with (15) and the monotonicity of the KL inverse (Prop. 1), we obtain the desired result (92):

$$\sup_{\pi_t^t, \dots, \pi_{T-1}^t} R_t \leq (T-t) \cdot \mathbb{D}^{-1} \left(\frac{\bar{R}_t^{\perp \star}}{T-t} \mid \mathbb{I}(o_t; s_t) \right) \quad (95)$$

$$= \bar{R}_t(a_{0:t-1}). \quad (96)$$

■

APPENDIX B CHERNOFF-HOEFFDING BOUND

In our numerical examples (Sec. VII), we utilize a slightly tighter version of Hoeffding's inequality than the one presented in Theorem 3. In particular, we use the following Chernoff-Hoeffding inequality (see [34, Theorem 5.1]).

Theorem 5 (Chernoff-Hoeffding inequality [34]). *Let z be a random variable bounded within $[0, 1]$, and let $z_1, \dots, z_n$ denote i.i.d. samples. Then, with probability at least $1 - \delta$ (over the sampling of $z_1, \dots, z_n$), the following bound holds with probability at least $1 - \delta$:*

$$\mathbb{D}_{\mathcal{B}} \left(\frac{1}{n} \sum_{i=1}^n z_i \parallel \mathbb{E}[z] \right) \leq \frac{\log(2/\delta)}{n}. \quad (97)$$

We can obtain an upper bound on $\mathbb{E}[z]$ using (97) as follows:

$$\mathbb{E}[z] \leq \sup \left\{ p \in [0, 1] \mid \mathbb{D}_{\mathcal{B}} \left(\frac{1}{n} \sum_{i=1}^n z_i \parallel p \right) \leq \frac{\log(2/\delta)}{n} \right\}. \quad (98)$$

The optimization problem in the RHS of (98) is analogous to the KL inverse defined in Sec. III-B, and can be thought of as a “right” KL inverse (instead of a “left” KL inverse). Similar to the KL inverse in Sec. III-B, we can solve the optimization problem in (98) using geometric programming.