Active Uncertainty Reduction for Human-Robot Interaction: An Implicit Dual Control Approach

Haimin Hu[1] and Jaime F. Fisac[1]

Department of Electrical and Computer Engineering, Princeton University, USA,
{haiminh,jfisac}@princeton.edu

Abstract. The ability to accurately predict human behavior is central to the safety and efficiency of robot autonomy in interactive settings. Unfortunately, robots often lack access to key information on which these predictions may hinge, such as people's goals, attention, and willingness to cooperate. Dual control theory addresses this challenge by treating unknown parameters of a predictive model as stochastic hidden states and inferring their values at runtime using information gathered during system operation. While able to optimally and automatically trade off exploration and exploitation, dual control is computationally intractable for general interactive motion planning, mainly due to the fundamental coupling between robot trajectory optimization and human intent inference. In this paper, we present a novel algorithmic approach to enable active uncertainty reduction for interactive motion planning based on the implicit dual control paradigm. Our approach relies on sampling-based approximation of stochastic dynamic programming, leading to a model predictive control problem that can be readily solved by real-time gradient-based optimization methods. The resulting policy is shown to preserve the dual control effect for a broad class of predictive human models with both continuous and categorical uncertainty. The efficacy of our approach is demonstrated with simulated driving examples.

Keywords: human-robot interaction, dual control theory, stochastic MPC

1 Introduction

Computing robot plans that account for possible interactions with one or multiple humans is a challenging task, as the robotic system and human agents may have coupled dynamics, limited communication capabilities, and conflicting interests. To achieve safety and efficiency in human-robot interaction settings, the robot must competently predict and seamlessly adapt to human behavior. Intent-driven behavior models are widely used for such predictions: for example, the Boltzmann model of noisily rational human decision-making [1, 2] assumes that the human is exponentially likelier to take actions with a higher underlying *utility*. If the human's intent is well captured by a given utility function, the interaction can be modeled as a dynamic game in which the human's and the robot's feedback strategies can be obtained via dynamic programming [3]. However, typical interaction settings may admit a plethora of *a priori* plausible intents (e.g., corresponding to distinct equilibrium solutions [4] or different human preferences [5]), which in general cannot be fully modeled, let alone observed, by the robot [6]. The

robot may seek to represent the human's intent through a parametric model and then infer the value of these parameters as hidden states under a Bayesian framework [6–8], but doing so tractably while planning through interactions is an open problem.

Multi-stage trajectory optimization with closed-loop Bayesian inference can be generally cast as a stochastic optimal control problem. An important aspect of stochastic control with hidden states is whether the computed policy generates the so-called *dual control* effect [9–11]; that is, in the context of human-robot interaction, whether the robot *actively* seeks to reduce the uncertainty about human's hidden states. Solution methods for dual stochastic optimal control problems can be categorized into *explicit* approaches [5, 12, 13], which reformulate the problem with some form of heuristic probing, and implicit approaches [10, 14, 15], which directly tackle the control problem with stochastic dynamic programming. While explicit dual control problems are in general easier to formulate and solve than their implicit counterparts, designing the probing term and tuning its weighting factor can be non-trivial and may lead to inconsistent performance. For a comprehensive review of dual control methods we refer to [11].

Contribution: In this paper, we formulate a broad class of interactive planning problems in the framework of stochastic optimal control and present an approximate solution method using implicit dual stochastic model predictive control (SMPC). The resulting policy automatically trades off the cost of exploration and exploitation, allowing the robot to actively reduce the uncertainty about the human's hidden states without sacrificing expected planning performance. Our proposed SMPC problem supports both continuous and categorical human uncertainty and can be solved using off-the-shelf real-time nonlinear optimization solvers. To our knowledge, this is the first human-robot interactive motion planning framework that performs active uncertainty reduction without requiring an explicit information gathering strategy or objective.

2   Related Work

Human-Robot Interaction as a POMDP. Robotic motion planning problems that involve identification of human intentions and behaviors can be modeled as a Mixed Observability Markov Decision Process [16], a variant of the well-known Partially Observable Markov Decision Process (POMDP). Although intractable in the general form, efficient algorithms such as DESPOT [17] and POMCP [18] have been developed to approximately solve POMDPs. In [19], a dual SMPC method is proposed to solve moderate-sized POMDPs. The resulting policy naturally exhibits system probing and automatically trades off exploration and exploitation. When interactions between the robot and the human are modeled, the POMDP formulation becomes a (usually intractable) Partially-Observable Stochastic Game (POSG) [5]. Our approach can be viewed as a new computationally efficient framework for solving human-robot interaction problems cast as POSGs.

Stochastic Model Predictive Control. SMPC has been widely used for robotic motion planning under uncertainty due to its ability to handle safety-critical constraints and general uncertainty models. In [20], an SMPC approach is proposed for lane change assistance of the ego vehicle in the presence of other human-driven vehicles. In [21], a scenario-based SMPC algorithm is proposed to capture multimodal reactive behaviors

of uncontrolled human agents. In [8], a provably safe SMPC planner is developed for human-aware robot planning, which proactively balances expected performance with the risk of high-cost emergency safety maneuvers triggered by low-probability human behaviors. However, all those SMPC methods do not produce dual control effect. In [15], an implicit dual SMPC is proposed for optimal control of nonlinear dynamical systems with both parametric and structural uncertainty. Our paper builds on this approach to enable active uncertainty reduction for human-robot interaction.

Active Information Gathering. To date, most human-in-the-loop planning methods follow a "passively adaptive" paradigm. In [4], human-robot interaction is modeled as a general-sum differential game with human's equilibrium uncertainty. The robot first infers which equilibrium the humans are operating at and then aligns its own strategy with the inferred human's equilibrium solution. Recently, the notion of active information gathering, which is conceptually very similar to dual control, has received attention from the human-robot interaction community. In [5], an additional information gathering term is added to the robot's nominal objective in a trajectory optimization framework for online estimating unknown parameters of the human's state-action value function. In fact, according to the categorization proposed by [11], the planning formulation in [5] can be classified as an explicit dual control approach, which requires heuristic design of a probing mechanism and weighing the relative importance between optimizing the expected performance and reducing the uncertainty of human's unknown parameters. On the contrary, we propose an implicit dual control approach, which automatically balances performance with uncertainty reduction, thus not requiring design of a heuristic information gathering mechanism.

3   Preliminaries

Human-Robot Joint System. We consider a class of discrete-time input-affine dynamics that capture the interaction between a human individual *or group* and a robotic system,

$$x_{t+1} = f(x_t) + B^R(x_t)u_t^R + B^H(x_t)u_t^H + d_t, \tag{1}$$

where $x_t \in \mathbb{R}^n$ is the joint state vector, $u_t^R \in \mathcal{U}^R \subseteq \mathbb{R}^{m_R}$ and $u_t^H \in \mathcal{U}^H \subseteq \mathbb{R}^{m_H}$ are the control vectors of the robot and the human, respectively, $f : \mathbb{R}^n \to \mathbb{R}^n$ is a nonlinear function that describes the autonomous part of the dynamics, $B^R : \mathbb{R}^n \to \mathbb{R}^{n \times m_R}$ and $B^H : \mathbb{R}^n \to \mathbb{R}^{n \times m_H}$ are control input matrices that can depend on the state, and d_t is an additive uncertainty term representing external disturbance inputs (e.g. wind) and modeling error. For simplicity, we further assume that $d_t \sim \mathcal{N}(0, \Sigma_d)$ is a zero-mean i.i.d. Gaussian random variable.

Human Action Parameterization. In this paper, we parametrize the human's action at each time t as a stochastic policy:

$$u_t^H := \sum_{i=1}^{n_\theta} \theta_i^M u_i^M, \quad \text{subject to} \quad u_t^H \in \mathcal{U}^H, \tag{2}$$

which is a linear combination of *stochastic basis policies* u_i^M with parameter $\theta^M := (\theta_1^M, \theta_2^M, \ldots, \theta_{n_\theta}^M) \in \mathbb{R}^{n_\theta}$. We further allow each basis policy u_i^M to have different *modes* M that take values from a *finite* set $\mathcal{M}$. We define each basis policy u_i^M with the
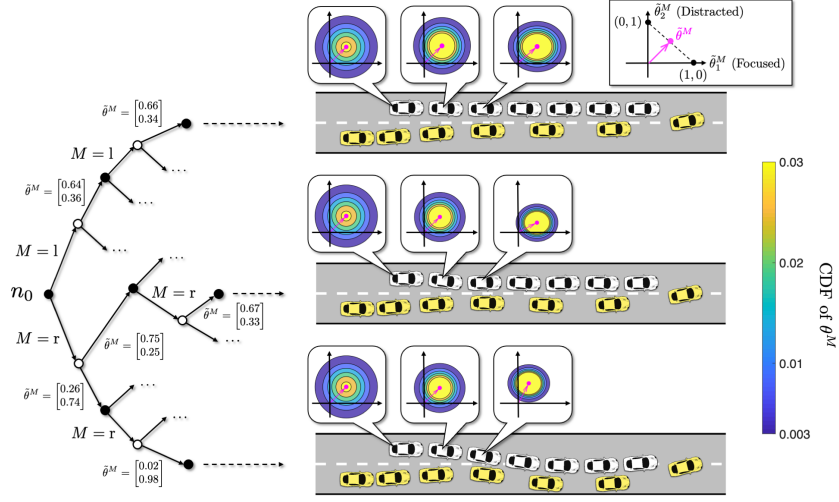
Fig. 1: Illustration of a scenario tree with $N^d = 2$ dual control time steps and $N^e = 4$ exploitation steps. The human-driven vehicle and autonomous car are plotted in white and yellow, respectively. The hidden state θ^M is modeled as a 2D Gaussian random variable, which is then projected onto a 1-simplex by a softmax operation. The contour plots display level sets of the cumulative distribution function (CDF) of θ^M. Uncertainty is *less* significantly reduced in the case where the human prefers the left lane (upper branch), since their behavior is less influenced by the robot's (probing) actions.

"noisily rational" Boltzmann model from cognitive science [1]. Under this model, the human picks actions according to a probability distribution:

$$u_i^M \sim p\left(u_i^M \mid x, u^R; M\right) = \frac{e^{-Q_i^M\left(x, u^R, u_i^M\right)}}{\int_{\tilde{u}_i^M \in \mathcal{U}^H} e^{-Q_i^M\left(x, u^R, \tilde{u}_i^M\right)} d\tilde{u}_i^M}, \tag{3}$$

for all $i = 1, 2, \ldots, n_\theta$ and $M \in \mathcal{M}$. Here, $Q_i^M(x, u^R, u_i^M)$ is the human's state-action value (or Q-value) function associated with the i-th basis function and mode M. This model assumes that, for a pair of fixed (i, M), the human is exponentially likelier to pick an action that maximizes the state-action value function.

Remark 1. Our approach is agnostic to the concrete methods for determining the human's state-action value function $Q_i^M(x, u^R, u_i^M)$, model parameter θ^M and mode M, which are usually specified by the system designer based on domain knowledge or learned from prior data. Goal-driven models of human motion are well-established in the literature. See for example [2, 5]. We provide two examples below.

Example 1. Consider the highway driving scenario depicted in Fig. 1, involving an autonomous car (R) and a human-driven vehicle (H), whose action is parameterized as

$$u_t^H = \theta_D^M \pi_D^M\left(x_t, u_t^R\right) + \theta_F^M \pi_F^M\left(x_t, u_t^R\right), \quad \text{subject to} \quad u_t^H \in \mathcal{U}^H,$$

where θ_D^M and θ_F^M capture the level of distraction and focus of the human, respectively. A focused human accounts for the safety of the joint system (e.g. avoiding and making

room for the robot when it attempts to merge in front of the human), while a distracted human does not. Modeling the human's level of distraction as a *continuum* is motivated by recent work [22], which differs from ours in that they assume an additive structure in the human's reward (objective) function rather than their policy. Nonetheless, per the maxim of "all models are wrong, but some are useful" [23], we will show in Section 6 that our choice of human predictive model is useful for planning purposes and leads to a policy that outperforms the state-of-the-art baseline. Further, the human has two distinct modes, namely preferring to drive in the left lane or in the right lane, i.e., $M \in \{\mathrm{l}, \mathrm{r}\}$. Given a joint state x_t, for each i and M we compute the human's game-theoretic state-action value function using the iterative linear-quadratic game method [24], which yields an approximate local feedback Nash equilibrium solution.

Example 2. In the second example, we use parameter $\theta^M := (\theta_C^M, \theta_{NC}^M)$, which captures the human's level of cooperativeness. A cooperative (C) human also optimizes for the robot's objective while a non-cooperative (NC) human does not. We define the discrete modes as different interactive behaviors of the human toward the robot, similar to [7, 16]. Specifically, the human's state-action value function is defined as

$$
Q_i^M = \begin{cases}
Q_i^M \left(x_t, u_t^{R,\mathrm{Nash}}(x_t), u_i^M \right), & \text{if } M = \mathrm{N} \text{ (Nash)}, \\
Q_i^M \left(x_t, u_t^{R,\mathrm{worst}}(x_t), u_i^M \right), & \text{if } M = \mathrm{p} \text{ (protected)}, \\
Q_i^M \left(x_t, u_t^{R,\mathrm{best}}(x_t), u_i^M \right), & \text{if } M = \mathrm{w} \text{ (wishful)}, \\
Q_i^M \left(x_t^H, u_i^M \right), & \text{if } M = \mathrm{o} \text{ (oblivious)}.
\end{cases}
$$

In the first three modes, the human assumes that the robot's control u_t^R is a local feedback Nash equilibrium solution [24], the worst-case one that minimizes Q_i^M, and the best-case one that maximizes Q_i^M, respectively; the last mode follows the same assumption as in [6, 8] that the human ignores the presence of other agents.

Inferring Model Parameter and Mode. In general, parameter θ^M and mode M in human action model (2) are *hidden states* that are unknown to the robot. Therefore, they can only be inferred from past observations. To address this, we define the information vector $\mathcal{I}_t := \left[x_t, u_{t-1}^R, \mathcal{I}_{t-1} \right]$ as the collection of all information that is *causally observable* by the robot at time $t \geq 0$, with $\mathcal{I}_0 = [x_0]$. We then define the *belief state* $b_t := p \left(\theta^M, M \mid \mathcal{I}_t \right)$ as the joint distribution of (θ^M, M) conditioned on $\mathcal{I}_t$, and $b_0 := p \left(\theta^M, M \right)$ is a given prior distribution. When the robot receives a new observation $x_{t+1} \in \mathcal{I}_{t+1}$, the current belief state b_t is updated using the recursive Bayesian inference:

$$
p(\theta_-^M \mid \mathcal{I}_{t+1}; M) = \frac{p(x_{t+1} \mid u_t^R, \mathcal{I}_t; \theta^M, M) p(\theta^M \mid \mathcal{I}_t; M)}{p(x_{t+1} \mid u_t^R, \mathcal{I}_t; M)}, \tag{4a}
$$

$$
p(M_- \mid \mathcal{I}_{t+1}) = \frac{p(x_{t+1} \mid u_t^R, \mathcal{I}_t; M) p(M \mid \mathcal{I}_t)}{p(x_{t+1} \mid u_t^R, \mathcal{I}_t)}, \tag{4b}
$$

$$
b_{t+1}^- := p(\theta_-^M, M_- \mid \mathcal{I}_{t+1}) = p(\theta_-^M \mid \mathcal{I}_{t+1}; M) p(M_- \mid \mathcal{I}_{t+1}), \tag{4c}
$$

$$
b_{t+1} = g^t(b_{t+1}^-) := \int p(\theta^M, M \mid \tilde{\theta}_-^M, \tilde{M}_-) p(\tilde{\theta}_-^M, \tilde{M}_- \mid \mathcal{I}_{t+1}) d\tilde{\theta}_-^M d\tilde{M}_-. \tag{4d}
$$

where $p(\theta^M, M \mid \tilde{\theta}^M_-, \tilde{M}_-)$ is a transition model and $g^t(\cdot)$ is the belief state transition dynamics. Compactly, we can rewrite (4a)-(4d) as a dynamical system,

$$b_{t+1} = g(b_t, x_{t+1}, u_t^R). \tag{5}$$

Unfortunately, system (5) in general does not adopt an analytical form beyond one-step evolution. Even if the prior distribution b_t is a Gaussian, the posterior b_{t+1} ceases to be a Gaussian since the human's action u_t^H defined by (2) and (3) (which in turn affects the observation x_{t+1}) is generally non-Gaussian, thus precluding the use of the conjugate-prior properties of Gaussian distributions. In Section 5, we will introduce a computationally efficient method to propagate the belief state dynamics approximately.

4   Problem Statement

Canonical Human-in-the-Loop Planning Problem. We now define the central problem we want to solve in this paper: the canonical human-in-the-loop planning problem, which is formulated as a stochastic finite-horizon optimal control problem as follows:

$$\min_{\pi_{[0:N-1]}^R} \mathbb{E}_{\substack{(\theta^M, M) \sim b_{[0:N-1]}, \\ u_{[0:N-1]}^H \sim (2), d_{[0:N-1]}}} \sum_{k=0}^{N-1} \ell^R \left(x_k, \pi_k^R(x_k, b_k) \right) + \ell_F^R(x_N) \tag{6a}$$

$$\text{s.t.}\quad x_0 = \hat{x}_t,\ b_0 = \hat{b}_t, \tag{6b}$$

$$x_{k+1} = f(x_k) + B^R \pi_k^R(x_k, b_k) + B^H u_k^H + d_k, \quad \forall k = 0, \dots, N-1 \tag{6c}$$

$$b_{k+1} = g\left(b_k, x_{k+1}, \pi_k^R(x_k, b_k) \right), \qquad\qquad \forall k = 0, \dots, N-1 \tag{6d}$$

$$x_k \notin \mathcal{F}, \qquad\qquad \forall k = 0, \dots, N \tag{6e}$$

where $\hat{x}_t$ and $\hat{b}_t$ are the state measured and belief state maintained at time t, $\ell^R : \mathbb{R}^n \times \mathcal{U}^R \to \mathbb{R}_{\geq 0}$ and $\ell_F^R : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ are designer-specified stage and terminal cost function, $\pi_k(x_k, b_k)$ is a *causal* feedback policy [10, 11] that leverages the (yet-to-be-acquired) knowledge of future state x_k and belief state b_k, and $\mathcal{F} \subseteq \mathbb{R}^n$ is a failure set that the state is not allowed to enter.

In theory, problem (6) can be solved using stochastic dynamic programming [25]. An optimal robot's value function $V_k(x_k, b_k)$ and control policy $\pi_k^{R,*}(x_k, b_k)$ can be obtained backwards in time using the Bellman recursion,

$$V_k(x_k, b_k) = \min_{\pi_k(x_k, b_k)} \ell^R(x_k, u_k^R) + \mathbb{E}_{\substack{(\theta^M, M) \sim b_k, \\ u_k^H \sim (2), d_k}} \left[V_{k+1}(x_{k+1}, b_{k+1}) \mid \mathcal{I}_k \right] \tag{7}$$

$$\text{s.t.}\ (6c) - (6e)$$

with terminal condition $V_N(x_N, b_N) = \ell_F^R(x_N)$.

Dual Control Effect. Value function $V_t(x_t, b_t)$ obtained by solving (7) depends on future belief states b_k $(k > t)$, thus giving the optimal policy $u_t^{R,*} := \pi_0^R(\hat{x}_t, \hat{b}_t)$ the ability to affect *future uncertainty* of the human quantified by the belief states. Therefore, the optimal policy $u_t^{R,*}$ of (7) possesses the property of *dual control effect*, defined

formally in Definition 1. Due to the principle of optimality [25], the policy achieves an optimal balance between optimizing the robot's expected performance objective (6a) and actively reducing its uncertainty about the human. In other words, the optimal policy of (7) automatically probes the human agents to reduce their associated uncertainty *only* to the extent that doing so improves the robot's expected closed-loop performance.

Definition 1 (**Dual Control Effect [9–11, 26]**). *A control input has dual control effect if it can affect, with nonzero probability, (a) at least one rth-order ($r \geq 2$) central moment of a hidden state variable in a metric space, or (b) the entropy of a categorical hidden state variable.*

Complication and Approximation of Dual Control. Unfortunately, (7) is computationally intractable in all but the simplest cases, mainly due to nested optimization of robot's action and and computing the conditional expectation. The expectation term in (7) can be approximated to arbitrary accuracy with quantization of the belief states, which, however, leads to exponential growth in computation, i.e. the issue of *curse of dimensionality* [25]. It is for those reasons that approximate methods are mainly used to solve dual control problems. Approximate dual control can be categorized into: explicit approaches, e.g. [5, 12, 13] that simplify the original stochastic optimal control problem by artificially introducing probing effect or information gathering objectives to the control policy, and implicit approaches, e.g. [10, 14, 15] that rely on direct approximation of the Bellman recursion (7). The approach we take in this paper is a *scenario-based implicit dual control method*, which is detailed in the next section. The main advantage of using the implicit dual control approximation is that the automatic exploration-exploitation trade-off of the policy is naturally preserved in an optimal sense [11, 19].

5   Active Uncertainty Reduction for Human-Robot Interaction

In this section, we describe an implicit dual control approach towards approximately solving the canonical human-in-the-loop planning problem (6). We start by presenting two approximation schemes for propagating the belief state dynamics and computing the expectation in Bellman recursion (7), which are necessary for reformulating (7) as a real-time solvable SMPC problem. The formulation and properties of our proposed SMPC problem are shown towards the end of this section. See Fig. 1 for an example solution of our proposed SMPC problem.

Tractable Reformulation of Belief State Dynamics. We start by deriving a tractable recursive update rule for the belief state dynamics (5) by approximating the human's action model (2), for a given (M, θ^M), as a Gaussian distribution. The technical tool we rely on is the Laplace approximation [27]. Precisely, the conditional probability distribution of each human's basis policy u_i^M defined in (3) is approximated as:

$$u_i^M \sim p\left(u_i^M \mid x_t, u_t^R; M\right) \approx \mathcal{N}\left(\mu_i^M(x_t, u_t^R), \Sigma_i^M(x_t, u_t^R)\right), \qquad (8)$$

where the mean function is $\mu_i^M(x_t, u_t^R) = \arg\max_{u_i^M \in \mathcal{U}^H} Q_i^M\left(x_t, u_t^R, u_i^M\right)$ and the covariance function is $\Sigma_i^M(x_t, u_t^R) = \nabla_{u_i^M}^2 \left. Q_i^M\left(x_t, u_t^R, u_i^M\right)^{-1}\right|_{u_i^M = \mu_i^M(x_t, u_t^R)}$. The

intuition behind the above Laplace approximation scheme is that the Gaussian distribution obtained in (8) centers around the mode of the original basis policy distribution $p\left(u_i^M \mid x_t, u_t^R, M\right)$, i.e. $\mu_i^M(x_t, u_t^R)$, which corresponds to the *perfectly rational* human action associated with θ_i^M. Hence, the overall human action distribution, conditioned on θ^M and M, is given by:

$$u_t^M \sim \mathcal{N}\left(\sum_{i=1}^{n_\theta} \theta_i^M \mu_i^M(x_t, u_t^R), \sum_{i=1}^{n_\theta} \left(\theta_i^M\right)^2 \Sigma_i^M(x_t, u_t^R)\right). \tag{9}$$

We subsequently lift the requirement that $u_t^H \in \mathcal{U}^H$ in order to keep u_t^H normally distributed *during belief propagation*, and we use projected u_t^H for state evolution. Now, we can rewrite the human-robot joint system (1) as

$$x_{t+1} = B^H(x_t) U^M(x_t, u_t^R)\theta^M + f(x_t) + B^R(x_t)u_t^R + \bar{d}_t^M, \tag{10}$$

where $U^M(x_t, u_t^R) := \left[\mu_1^M(x_t, u_t^R)\ \mu_2^M(x_t, u_t^R)\ \ldots\ \mu_{n_\theta}^M(x_t, u_t^R)\right]$ is a deterministic state- and input-dependent matrix, and the combined disturbance term $\bar{d}_t^M$ is a zero-mean Gaussian, whose covariance is given by

$$\Sigma^{\bar{d}_t^M}(x_t, u_t^R; \theta^M) := \Sigma_d + \sum_{i=1}^{n_\theta} \left(\theta_i^M\right)^2 B^H(x_t)\Sigma_i^M(x_t, u_t^R)B^H(x_t)^\top. \tag{11}$$

Note that even if (10) is linear in θ^M, dependence of covariance matrix $\Sigma^{\bar{d}_t^M}$ on θ^M, as shown in (11), still prohibits updating the belief states in closed-form. To this end, we approximate covariance $\Sigma^{\bar{d}_t^M}$ by fixing θ^M with some estimated value $\bar{\theta}^M$, which can be obtained, for example, from solutions of the last run or roll-out-based simulations.

Now, we can readily propagate approximate belief state dynamics (5) efficiently in closed-form, leveraging the conjugate prior property of Gaussian distributions. Crucially, the above approximate belief state updates preserve the dual control effect. To verify this claim, we examine the covariance of the updated conditional distribution $p(\theta_-^M \mid \mathcal{I}_{t+1}; M)$:

$$\Sigma_{t+1}^{\theta_-^M} = \Sigma_t^{\theta^M} + \left[\left(B^H U^M(x_t, u_t^R)\right)^\top \Sigma^{\bar{d}_t^{M^{-1}}}(x_t, u_t^R)B^H U^M(x_t, u_t^R)\right]^{-1}, \tag{12}$$

whose prior distribution is $p(\theta^M \mid \mathcal{I}_t; M) = \mathcal{N}(\mu_t^{\theta^M}, \Sigma_t^{\theta^M})$. From the above we can clearly see that the robot's control u_t^R affects the updated covariance matrix $\Sigma_{t+1}^{\theta_-^M}$, and all future covariance matrices implicitly by affecting future states x_k $(k > t+1)$ via the Bellman recursion, hence producing dual control effect for θ^M according to Definition 1. Dual control effect for M can be similarly verified via (4b). We denote the approximate belief state dynamics as $b_{t+1} = \tilde{g}(b_t, x_{t+1}, u_t^R)$.

Implicit Dual Control using Scenario Trees. In this section, we propose an approximate solution method for the canonical human-in-the-loop planning problem (6) using scenario-tree-based stochastic model predictive control (ST-SMPC) [11, 28], which yields a control policy with dual control effect. The key idea of ST-SMPC is to approximate the expectation in Bellman recursion (7) based on uncertainty samples, i.e. quantized belief states. This leads to a scenario tree that allows us to roll out (7) as a deterministic finite-horizon optimal control problem, which can be readily solved by

gradient-based algorithms. Since our approach hinges on directly approximating Bell-man recursion (7), it can be understood as an implicit dual control method [11].

We denote a *node* in the scenario tree as n, whose time, state and belief state are denoted as t_n, x_n and b_n, respectively. Similarly, the uncertainty samples of the node are θ_n^M, $\bar{d}_n^M$ and M_n. Here, recall that $\bar{d}_n^M$ is the combined disturbance defined in (11), which implicitly incorporates a sample of the human's action. The set of all nodes is defined as $\mathcal{N}$. We define the transition probability from a parent node $\mathfrak{p}(n)$ to its child node n as $\bar{P}_n := P(\theta_n^M \mid \mathcal{I}_{\mathfrak{p}(n)}; M_n)P(\bar{d}_n^M \mid \mathcal{I}_{\mathfrak{p}(n)}; M_n)P\left(M_n \mid \mathcal{I}_{\mathfrak{p}(n)}\right)$. Subsequently, the *path transition probability* of node n, i.e. the transition probability from the root node n_0 to node n can be computed recursively as $P_n := \bar{P}_n \cdot \bar{P}_{\mathfrak{p}(n)} \cdots \bar{P}_{n_0}$. In order to quickly compute the conditional probabilities of $(\theta_n^M, \bar{d}_n^M)$ and avoid online sampling (i.e. during the optimization), we use an offline sampling procedure, leveraging the fact that they are Gaussian random variables, similar to what is done in [29]. We first generate samples offline from the standard Gaussian distribution. Then, during online optimization, these samples are transformed using the analytical mean and covariance:

$$\theta_n^M = \mu^{\theta_n^M}\left(x_n, u_n^R\right) + \left(\Sigma^{\theta_n^M}\left(x_n, u_n^R\right)\right)^{1/2} \theta_n^{M,o}, \tag{13a}$$

$$\bar{d}_n^M = \left(\tilde{\Sigma}^{\bar{d}_n^M}\left(x_n, u_n^R\right)\right)^{1/2} \bar{d}_n^{M,o}. \tag{13b}$$

Unlike existing ST-SMPC methods for human-robot interaction, such as [8, 20], our scenario tree has both *state- and input-dependent* uncertainty realizations (via transformations (13a) and (13b)) and path transition probabilities (via Bayesian update of M in (4b)). Therefore, the nodes can move in response to predicted states and inputs during online optimization. Fundamentally, it is this feature that allows the robot to actively reduce the human-related uncertainty via the dual control effect, and capture mutual responses between the human and the robot.

Remark 2. In order to alleviate the exponential growth of complexity associated with the scenario tree, only a small set of M, $\theta_n^{M,o}$ and $\bar{d}_n^{M,o}$ are sampled. A scenario pruning mechanism is introduced in [8] for static scenario trees, which can be used here as a heuristic to prune branches based on initial guesses. Principled pruning for scenario trees involving state- and input-dependent samples, however, remains an open question.

Exploitation Steps. In order to alleviate the computation challenge caused by the exponential growth of nodes in the scenario tree, we can stop branching the tree at a stage $N^d < N$, which we refer to as the dual control horizon. Subsequently, the remaining $N^e := N - N^d$ stages become the exploitation horizon, where each scenario is extended without branching and the belief states are only propagated with the transition dynamics $g^t(\cdot)$ defined in (4d), corresponding to a non-dual SMPC problem. Thanks to the scenario tree structure, control inputs of the exploitation steps still preserve the causal feedback property, allowing the robot to be cautious and "passively adaptive" to future human's uncertainty realizations.

Overall ST-SMPC Problem. Given a scenario tree defined by node sets $\mathcal{N}_t$, we can roll out the Bellman recursion (7) based on uncertainty samples in the tree, leading to a

ST-SMPC problem formulated as

$$\min_{\mathbf{U}_t^R} \sum_{\tilde{n} \in \mathcal{N}_t \setminus \mathcal{L}_t} P_{\tilde{n}} \ell^R(x_{\tilde{n}}, u_{\tilde{n}}^R) + \sum_{\tilde{n} \in \mathcal{L}_t} P_{\tilde{n}} \ell_F^R(x_{\tilde{n}}, b_{\tilde{n}})$$

$$\begin{aligned}
\text{s.t. } & x_{n_0} = \hat{x}_t, \ b_{n_0} = \hat{b}_t, \\
& x_{\tilde{n}} = f(x_{\mathfrak{p}(\tilde{n})}) + B^R(x_{\mathfrak{p}(\tilde{n})})u_{\mathfrak{p}(\tilde{n})}^R + B^H(x_{\mathfrak{p}(\tilde{n})})u_{\tilde{n}}^H + \bar{d}_{\tilde{n}}^M, && \forall \tilde{n} \in \mathcal{N}_t \setminus \{n_0\} \\
& b_{\tilde{n}} = \tilde{g}\left(b_{\mathfrak{p}(\tilde{n})}, x_{\tilde{n}}, u_{\mathfrak{p}(\tilde{n})}^R\right), && \forall \tilde{n} \in \mathcal{N}_t^d \\
& b_{\tilde{n}} = g^t\left(b_{\mathfrak{p}(\tilde{n})}\right), && \forall \tilde{n} \in \mathcal{N}_t^e \\
& u_{\tilde{n}}^H = U^{M_{\tilde{n}}}(x_{\mathfrak{p}(\tilde{n})}, u_{\mathfrak{p}(\tilde{n})}^R)\theta_{\tilde{n}}^M, \ u_{\tilde{n}}^H \in \mathcal{U}^H, \text{ (13a), (13b)}, && \forall \tilde{n} \in \mathcal{N}_t \setminus \{n_0\} \\
& u_{\tilde{n}}^R \in \mathcal{U}^R, && \forall \tilde{n} \in \mathcal{N}_t \setminus \mathcal{L}_t \\
& x_{\tilde{n}} \notin \mathcal{F}, && \forall \tilde{n} \in \mathcal{N}_t
\end{aligned}$$
(14)

where $\mathcal{L}_t$ is the set of all leaf nodes, (i.e. ones that do not have a descendant), $\mathcal{N}_t^d$ and $\mathcal{N}_t^e$ are the set of dual control and exploitation nodes, respectively, and set $\mathbf{U}_t^R := \{u_{\tilde{n}}^R \in \mathbb{R}^{m_R} : \tilde{n} \in \mathcal{N}_t \setminus \mathcal{L}_t\}$ is the collection of robot's control inputs associated with all non-leaf nodes. Problem (14) is a nonconvex trajectory optimization problem, which can be solved using general-purpose nonconvex solvers such as SNOPT [30]. The optimal solution $\mathbf{U}_t^{R,*}$ to (14) is implemented in a receding horizon fashion, i.e. $\pi_{\text{ID-SMPC}}^R(\hat{x}_t, \hat{b}_t) := u_{n_0}^{R,*}$.

Theorem 1. *The policy $\pi_{ID\text{-}SMPC}^R(\cdot, \cdot)$ obtained by solving (14) has dual control effect.*

Proof. From (12), the robot's control u_t^R can affect the covariance of the belief over θ^M. Therefore, the policy $\pi_{\text{ID-SMPC}}^R$ has dual control effect for θ^M per Definition 1. Similarly, from (4b), it can be seen that u_t^R can affect all components of the categorical distribution over M, and thereby its entropy, implying dual control effect for M per Definition 1.

Safety and Feasibility. Since the ST-SMPC problem (14) is formulated as a standard scenario program, safety assurances can be obtained based on existing safe control methods. In the following, we briefly discuss three ways to enforce safety, i.e. the human-robot joint system in closed-loop with the policy of (14) satisfying $x \notin \mathcal{F}$.

- *Robust.* Our prior work [8] provides robust safety and recursive feasibility guarantees for ST-SMPC based on shielding, a "least-restrictive" supervisory safety filter designed purely based on sets $\mathcal{F}, \mathcal{U}^H, \mathcal{U}^R$, and is agnostic to the specific predictive model of human motion. As a special case, [29] provides closed-loop safety guarantees by projecting the ST-SMPC policy onto the set of backup control policies associated with a robust controlled-invariant set. Note that an additional assumption of bounded disturbance is needed in order to obtain robust safety guarantees.
- *Probabilistic.* We can replace (6e) with *chance constraints* to obtain probabilistic safety guarantees, i.e. $P[x \in \mathcal{F}] \leq \beta$, where β is the tolerance level, provided that $x \in \mathcal{F}$ can be written as a set of inequality constraints. An analytical bound of β and probabilistic feasibility guarantee are established for chance-constrained ST-SMPC problems with dynamic obstacles in [31].
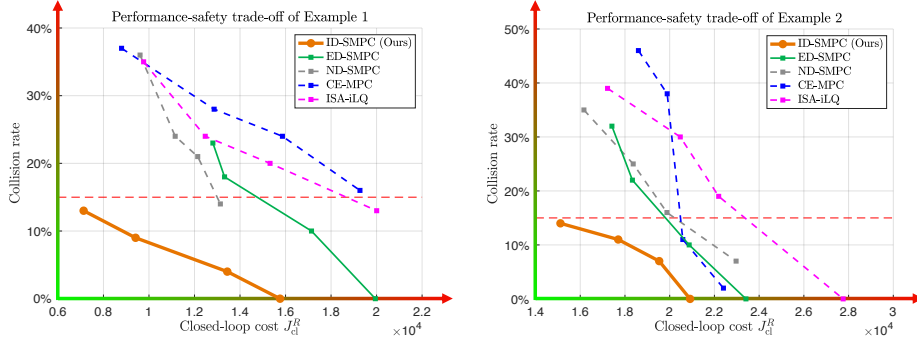
Fig. 2: Performance and safety trade-off of Examples 1 and 2. Each data point is obtained based on a particular design of the planner and the statistical data of 100 trials with different random seeds. The collision rate of ID-SMPC is always below 15% even for the least conservative design.

- *Soft constrained.* Soft constraint is a simple, yet effective technique widely used by MPC to enforce safety of the closed-loop system. The advantages of soft constraints are that they are easy to implement, efficient to optimize, and feasibility can be guaranteed. In this paper, we tailor the soft constrained MPC approach in [32] for ST-SMPC, which relaxes the original hard constraints $x_{\tilde{n}} \notin \mathcal{F}$ in (14) with slack variables for each node $\tilde{n} \in \mathcal{N}_t$. Although satisfaction of $x \notin \mathcal{F}$ can no longer be guaranteed for the closed-loop system, we show in Section 6 that our approach significantly reduces the collision rate comparing to the baselines (using the same soft constrained MPC approach) due to active reduction of the human's uncertainty.

6   Simulation Results

We evaluate our proposed implicit dual scenario-tree-based SMPC (ID-SMPC) planner on simulated driving scenarios, where the human-driven vehicles are simulated using a game-theoretic model synthesized with [24]. The open-source code is available online.[1]

Baselines. We compare our proposed ID-SMPC planner against four baselines:

- Explicit dual SMPC (ED-SMPC) planner, which augments the stage cost ℓ^R in (6a) with an information gain term $\lambda(H(b_k) - H(b_{k+1}))$ proposed by [5]. Here, $\lambda > 0$ is a fine-tuned weighting factor.
- Non-dual scenario-tree-based SMPC (ND-SMPC) planner, which is based on solving (6) with a scenario tree that does not propagate belief states with an observation model (so the resulting policy does not have dual control effect). A similar scenario program is also considered in [8, 20, 28].
- Certainty-equivalent MPC (CE-MPC) planner, which is based on solving (6) with the certainty-equivalence principle [11, 15].

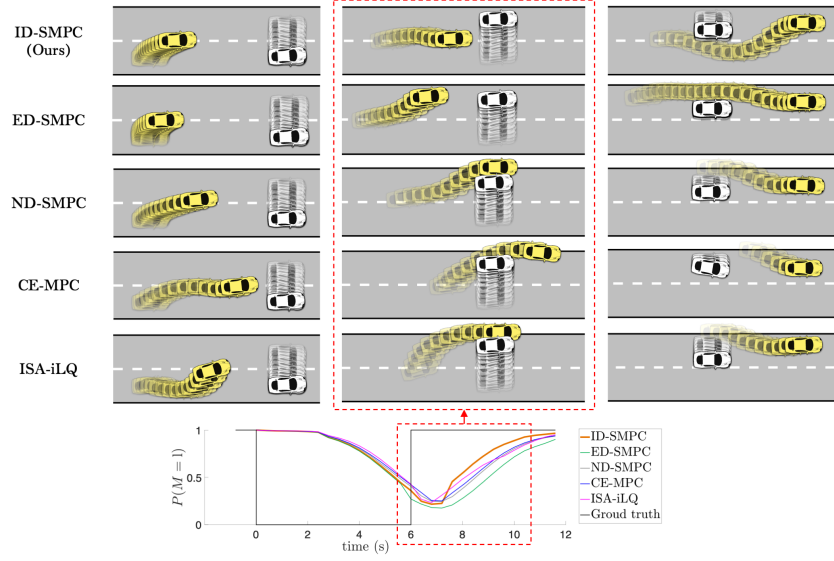[1] https://github.com/SafeRoboticsLab/Dual_Control_HRI

Fig. 3: Simulation snapshots of Example 1. Longitudinal positions are shown in relative coordinates with $p_x^H = 0$. The left, middle, and right columns display trajectories for $t = [0, 3]$ s, $t = [3, 5]$ s, and the remainder of the trajectories, respectively. The bottom figure shows probability $P(M = 1)$ for all four planners overtime. Our proposed ID-SMPC planner yielded a clean and sharp overtaking maneuver of the robot while the non-dual planners led to unsafe trajectories.

– Iterative linear-quadratic game with inference-based strategy alignment (ISA-iLQ) planner, proposed by [4] and originally developed in [24]. An input projector is used to enforce $u^H \in \mathcal{U}^H$ and $u^R \in \mathcal{U}^R$.

ED-SMPC is a dual control planner while the other three do not generate dual control effect. All planners use the same quadratic cost functions ℓ^R and ℓ_F^R.

Simulation Setup. All planners are equipped with the same human intent inference scheme. Vehicle and pedestrian dynamics are described by the kinematic bicycle model in [3] and the unicycle model in [24], respectively, both discretized with a time step of $\Delta t = 0.2$ s. The interactive human agents are simulated using the iterative linear-quadratic game method [24] with a game horizon of 3 s. All simulations are performed using MATLAB and YALMIP [33] on a desktop with an Intel Core i7-10700K CPU. All nonlinear MPC problems are solved with SNOPT [30]. The average solving time for ISA-iLQ, CE-MPC, ND-SMPC, ED-SMPC, and ID-SMPC at each time step are 1.92 s, 0.333 s, 0.184 s, 0.481 s, and 0.478 s, respectively, when the robot is interacting with one human agent (Section 6.1 and 6.2), and 19.3 s, 0.775 s, 1.37 s, 4.17 s, and 4.03 s, respectively, in the presence of three human agents (Section 6.3). The increase in solving time (for ST-SMPC problems) is due to the exponential growth of scenarios when the number of agents increases. This highlights an opportunity to improve time efficiency in multi-agent settings leveraging sparse scenario tree methods [8] and distributed MPC techniques.

ID-SMPC (safe and successful overtake)   CE-MPC (loss of safety)   ISA-iLQ (loss of liveness)
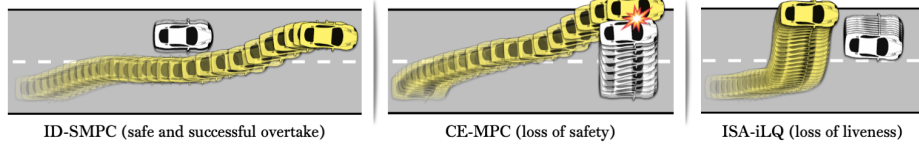
Fig. 4: One trial of Example 1, where the robot successfully overtook the human-driven vehicle in 6 s using ID-SMPC. CE-MPC and ISA-iLQ led to loss of safety and liveness, respectively.

Metrics. To measure the planning performance, we consider the following two metrics:

- Closed-loop cost, defined as $J_{\mathrm{cl}}^R := \sum_{t=0}^{T_{\mathrm{sim}}} \ell^R(x_t, u_t^R)$, where T_{sim} is the simulation horizon, and $x_{[0:T_{\mathrm{sim}}]}, u_{[0:T_{\mathrm{sim}}]}$ are the *executed* trajectories (with replanning).
- Collision rate, defined as $N_{\mathrm{coll}}/N_{\mathrm{trial}} \times 100\%$, where N_{coll} is the number of trials that a collision happens, i.e. $x_t \in \mathcal{F}$, and N_{trial} is the total number of trials.

Hypotheses. We make three hypotheses, which are supported by our simulation.

- **H1 (Performance and Safety Trade-off).** *Dual-control planners result in a better performance-safety trade-off than non-dual baselines.*
- **H2 (Safety and Liveness).** *Insufficient knowledge about human hidden states (due to lack of dual control effect) can cause loss of safety (e.g. collisions) and/or liveness (e.g. failing to overtake the human).*
- **H3 (Implicit vs Explicit Dual Control).** *Explicit dual control is less efficient than its implicit counterpart, even with fine tuning.*

6.1   Objective and Awareness Uncertainty (Example 1)

The performance-safety trade-off curve for Example 1 plotted in Fig. 2 validates H1. Here, we design each planner with a set of fine-tuned parameters, e.g. weight of the collision avoidance cost and acceleration limits. For a given planner design, we simulate the scenario 100 times, each with a different random seed, which affects three uncertainty sources: initial conditions, trajectories of the hidden states, and additive disturbances. Although the ED-SMPC policy also manages to achieve a low collision rate thanks to its ability to actively reduce the human uncertainty, its overall closed-loop performance is consistently inferior to that of ID-SMPC, which validates H3.

Trajectory snapshots and evolution of $P(M = 1)$ of one simulation trial are shown in Fig. 3. The priors are chosen as $P(M = 1) = 1$ and $\theta^{\mathrm{l}}, \theta^{\mathrm{r}} \sim \mathcal{N}((0.5, 0.5), 5I)$. Unlike non-dual control planners, ID-SMPC controlled the robot to approach the human-driven vehicle along the center of the road, allowing the robot to informatively probe the human—which resulted in a more accurate prediction of M (bottom)—and guiding the robot through a region from which collisions can be avoided more easily. Indeed, as the human's hidden state M switched from r to l at $t = 6$ s, the robot using ID-SMPC executed a sharp right turn and successfully avoided colliding with the human. The ED-SMPC planner, although effective at reducing the uncertainty at the beginning, failed to recognize that overtaking the human from the right would have resulted a more efficient trajectory. All non-dual control planners, *even with replanning*, caused a collision with
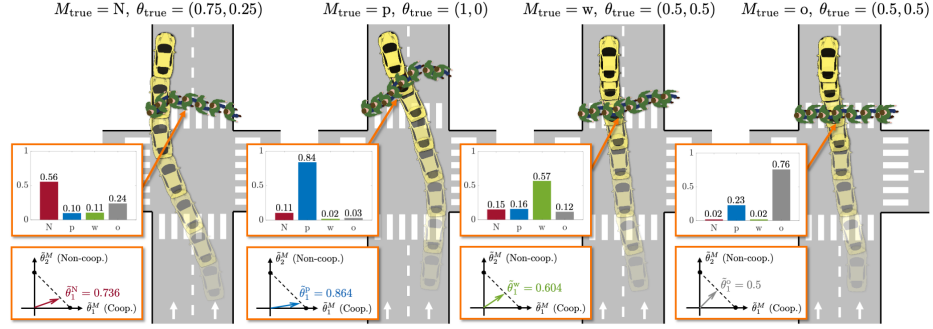
Fig. 5: Simulation snapshots of Example 2 using the proposed ID-SMPC planner. The ground-truth hidden states are fixed throughout the simulation. In the orange blocks we display the robot's running belief $p(M \mid \mathcal{I}_t)$ and maximum *a posteriori* mean of θ^M with $t = 2.6$ s. The robot was able to quickly identify the human's hidden states and planned a collision-free trajectory in all four trials, accounting for the anticipated uncertainty reduction and interactions with the human.

the human due to insufficient knowledge about M. It is also worth noticing that even if ID-SMPC uses ND-SMPC solutions for initialization, their closed-loop behaviors are vastly different, which essentially comes from the dual control effect.

In Fig. 4, we examine another simulation trial of Example 1. Ground truth values of the hidden states are $\theta = 1$ and $M = l$, respectively. Using ID-SMPC, the robot was able to safely overtake the human-driven car in 6 s. However, using the ISA-iLQ planner, the robot failed to overtake the human within 10 s, which is the simulation horizon. Due to lack of dual control effort, the robot was stuck behind the human, *unaware of the human's willingness to make room for the robot*. Fig. 4 also shows an unsafe trajectory generated with the CE-MPC planner. Results shown by Fig. 3 and Fig. 4 demonstrate that with dual control effort, the robot gains better safety and liveness when interacting with the human, which supports our hypothesis H2.

6.2   Behavioral and Cooperative Uncertainty (Example 2)

Next, we consider an uncontrolled traffic intersection scenario with the human uncertainty introduced in Example 2. Trajectory snapshots of four simulation trials with different hidden states are shown in Fig. 5. We chose uninformative prior distributions $p(M \mid \mathcal{I}_0) = \begin{bmatrix} 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix}$ and $\theta^M \sim \mathcal{N}((0.5, 0.5), 5I)$ for all $M \in \mathcal{M}$.

6.3   Multi-Agent Case Study

Finally, we apply ID-SMPC to the same intersection scenario as in Example 2 involving three human agents: two human-driven vehicles and a pedestrian. The uncertainty for the human-driven vehicles and the pedestrian is modeled with Example 1 and 2, respectively. Trajectory snapshots of one representative trial is shown in Fig. 6.
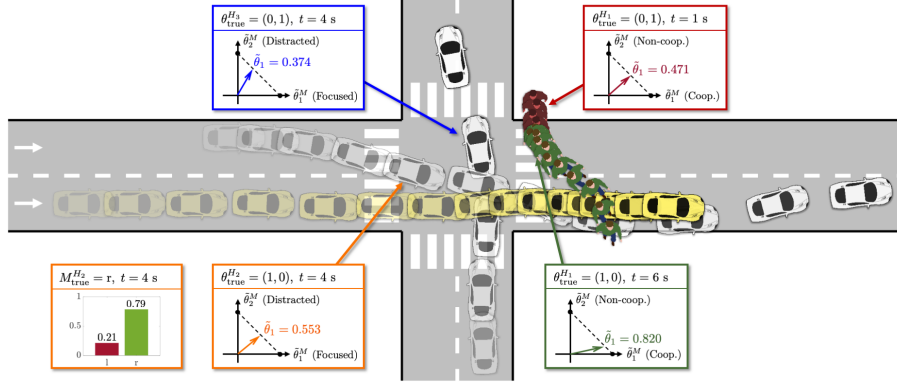
Fig. 6: Simulation snapshots and estimated hidden states of a multi-agent interaction scenario with a pedestrian (H_1) and two human-driven vehicles (H_2 and H_3) using the ID-SMPC planner.

7   Discussion and Future Work

We have introduced an implicit dual control approach towards active uncertainty reduction for human-robot interaction. The resulting policy improves safety and efficiency of human-in-the-loop motion planning via a tractable approximation to the Bellman recursion of a dual control problem, automatically achieving an efficient balance between improving expected performance and eliciting information on future human behavior. We see our work as an important step towards a broader class of methods that can handle general parametrizations of human behavior, including the effect of robot decisions on the human's hidden state. While this paper focuses on the robot's own performance, our approach may be adapted to account for social coordination and altruism [22] in cooperative human-robot settings. Finally, there is an open opportunity to derive formal probabilistic or worst-case safety guarantees based on results established in [8, 31].

References

[1]   R. D. Luce.   "Individual Choice Behavior". Oxford, England: John Wiley, 1959.

[2]   B. D. Ziebart et al.   "Maximum entropy inverse reinforcement learning." *AAAI*. 2008.

[3]   J. F. Fisac et al.   "Hierarchical game-theoretic planning for autonomous vehicles". *IEEE International Conference on Robotics and Automation (ICRA)*. 2019.

[4]   L. Peters et al.   "Inference-based strategy alignment for general-sum differential games". *19th International Conference on Autonomous Agents and Multi Agent Systems*. 2020.

[5]   D. Sadigh et al.   "Planning for cars that coordinate with people: leveraging effects on human actions for planning and active information gathering over human internal state". *Autonomous Robots* 42.7 (2018).

[6]   J. F. Fisac et al.   "Probabilistically Safe Robot Planning with Confidence-Based Human Predictions". *Robotics: Science and Systems*. 2018.

[7]   R. Tian et al.   "Safety Assurances for Human-Robot Interaction via Confidence-aware Game-theoretic Human Models". *CoRR* abs/2109.14700 (2021).

[8]   H. Hu, K. Nakamura, and J. F. Fisac.   "SHARP: Shielding-aware robust planning for safe and efficient human-robot interaction". *IEEE Robotics and Automation Letters* (2022).

[9]   A. A. Feldbaum. "Dual control theory. I". *Avtomatika i Telemekhanika* (1960).

[10] Y. Bar-Shalom and E. Tse. "Dual effect, certainty equivalence, and separation in stochastic control". *IEEE Trans. Autom. Control* 19.5 (1974).

[11] A. Mesbah. "Stochastic model predictive control with active uncertainty learning: A survey on dual control". *Annual Reviews in Control* 45 (2018).

[12] T. A. N. Heirung, B. Foss, and B. E. Ydstie. "MPC-based dual control with online experiment design". *Journal of Process Control* 32 (2015).

[13] R. Tian et al. "Anytime game-theoretic planning with active reasoning about humans' latent states for human-centered robots". *IEEE International Conference on Robotics and Automation (ICRA)*. 2021.

[14] E. D. Klenske and P. Hennig. "Dual control for approximate bayesian reinforcement learning". *The Journal of Machine Learning Research* 17.1 (2016).

[15] E. Arcari et al. "Dual stochastic MPC for systems with parametric and structural uncertainty". *Learning for Dynamics and Control*. 2020.

[16] T. Bandyopadhyay et al. "Intention-aware motion planning". *Algorithmic Foundations of Robotics X*. Springer, 2013.

[17] N. Ye et al. "DESPOT: Online POMDP Planning with Regularization". *Journal of Artificial Intelligence Research* 58.1 (2017).

[18] D. Silver and J. Veness. "Monte-Carlo planning in large POMDPs". *Neural Information Processing Systems*. 2010.

[19] M. A. Sehr and R. R. Bitmead. "Tractable dual optimal stochastic model predictive control: An example in healthcare". *IEEE Conference on Control Technology and Applications (CCTA)*. 2017.

[20] G. Schildbach and F. Borrelli. "Scenario model predictive control for lane change assistance on highways". *IEEE Intelligent Vehicles Symposium (IV)*. 2015.

[21] Y. Chen et al. "Interactive Multi-Modal Motion Planning With Branch Model Predictive Control". *IEEE Robotics and Automation Letters* 7.2 (2022).

[22] B. Toghi et al. "Cooperative autonomous vehicles that sympathize with human drivers". *IEEE Transactions on Intelligent Vehicles (IEEE TIV)* (2021).

[23] G. E. Box. "All models are wrong, but some are useful". *Robustness in Statistics* (1979).

[24] D. Fridovich-Keil et al. "Efficient iterative linear-quadratic approximations for nonlinear multi-player general-sum differential games". *IEEE International Conference on Robotics and Automation (ICRA)*. 2020.

[25] R. Bellman. "Dynamic programming". *Science* 153.3731 (1966).

[26] O. Hijab. "Entropy and dual control". *IEEE Conference on Decision and Control*. 1984.

[27] A. D. Dragan and S. S. Srinivasa. "Formalizing Assistive Teleoperation". *Robotics: Science and Systems VIII*. 2012.

[28] D. Bernardini and A. Bemporad. "Stabilizing model predictive control of stochastic constrained linear systems". *IEEE Trans. Autom. Control* 57.6 (2011).

[29] A. D. Bonzanini, J. A. Paulson, and A. Mesbah. "Safe learning-based model predictive control under state-and input-dependent uncertainty using scenario trees". *IEEE Conference on Decision and Control (CDC)*. 2020.

[30] P. E. Gill, W. Murray, and M. A. Saunders. "SNOPT: An SQP algorithm for large-scale constrained optimization". *SIAM Review* 47.1 (2005).

[31] O. de Groot et al. "Scenario-Based Trajectory Optimization in Uncertain Dynamic Environments". *IEEE Robotics and Automation Letters* 6.3 (2021).

[32] M. N. Zeilinger, M. Morari, and C. N. Jones. "Soft constrained model predictive control with robust stability guarantees". *IEEE Trans. Autom. Control* (2014).

[33] J. Löfberg. "YALMIP : A toolbox for modeling and optimization in MATLAB". *Proc. of the CACSD Conference*. Taipei, Taiwan, 2004.