2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Active Learning with Dual Model Predictive Path-Integral Control for Interaction-Aware Autonomous Highway On-ramp Merging

Jacob Knaup[1], Jovin D'sa[2], Behdad Chalaki[2], Tyler Naes[2],
Hossein Nourkhiz Mahjoub[2], Ehsan Moradi-Pari[2], Panagiotis Tsiotras[1]

***Abstract*— Merging into dense highway traffic for an autonomous vehicle is a complex decision-making task, wherein the vehicle must identify a potential gap and coordinate with surrounding human drivers, each of whom may exhibit diverse driving behaviors. Many existing methods consider other drivers to be dynamic obstacles and, as a result, they are incapable of capturing the full intent of the human drivers through this passive planning. In this paper, we propose a novel dual control framework based on Model Predictive Path-Integral control to generate interactive trajectories. This framework incorporates a Bayesian inference approach to *actively* learn the agents' parameters, i.e., other drivers' model parameters. The proposed framework employs a sampling-based approach that is suitable for real-time implementation through the utilization of GPUs. We illustrate the effectiveness of our proposed methodology through comprehensive numerical simulations conducted in both high and low-fidelity simulation scenarios focusing on autonomous on-ramp merging.**

I. Introduction

Autonomous driving, similar to many problems arising in robotics, requires optimal and safe decision-making in the presence of uncertainty about other agents' and humans' behavior [1]. Since every person may have their own unique driving style, an autonomous vehicle must learn a unique model for each other driver and then plan its trajectory accordingly in order to safely and efficiently interact with them [2]. The task of highway on-ramp merging has been shown to be an especially challenging task for autonomous vehicles and even human drivers [3], [4]. In congested traffic scenarios, the merging vehicle must successfully negotiate with other drivers to find someone who will yield to them and open a gap to merge into, while also dealing with constraints such as the end of the merge-lane [5].

There are different ways to approach these scenarios. For example, [6] uses an end-to-end learning-based approach in which deep reinforcement learning (RL) is used to learn an optimal policy for the merging vehicle. Meanwhile, [7] utilizes a stochastic model predictive control (SMPC) framework, wherein the authors formulate a scenario tree of possible traffic vehicle actions and solve the resulting nonlinear program (NLP). The authors of [8], [9] take a similar approach, but use Gaussian Process regression and a stochastic game theory approach, respectively, to model the unknown behavior of the other vehicles. Finally, references [10], [11] take a hierarchical approach in which a high-level RL algorithm is trained to provide a reward or policy to a low-level MPC algorithm used for trajectory generation.

Selecting a fully learning-based framework for autonomous agents' decision-making can raise concerns due to its lack of transparency and interpretability [12], especially in safety-critical scenarios. Thus, in this work, we propose an interpretable model-based stochastic optimal control framework. However, learning an *interaction-aware* model for the behavior of a given driver in highly interactive scenarios is a challenging problem on its own. One way to achieve this is to probe the driver by specific actions of the autonomous car and observe how they respond [13]. However, this involves a trade-off between exploring many potential yielding drivers and attempting to exploit the most promising drivers. Therefore, a *dual control paradigm* is extremely attractive in these scenarios [14]. In a dual control framework, the stochastic optimal control policy is designed with regards to not only the current uncertainty in the system model, but also it takes into account how the future actions of the autonomous vehicle and the surrounding drivers will provide new information to better inform its model. Thus, a dual control framework *actively learns* new information about other drivers while accomplishing the control objectives.

Due to the real-time execution requirements, since the problem of highway merging involves dynamic obstacles, and since an interaction-aware driver model is inherently nonlinear, the gradient-free Model Predictive Path-Integral control (MPPI) algorithm is an attractive choice for this problem [15]. MPPI relies on randomly sampling control sequences and evaluating the cost of the corresponding predicted trajectories to compute the optimal control sequence as a weighted average of the sampled actions.

Several stochastic variations of MPPI have been previously proposed in the literature. For example, in [16], the authors sample model parameter realizations in addition to control sequences to design a control sequence that is robust to potential model variations. Uncertainty-averse MPPI pairs MPPI with a mixture density network (a neural network trained to output the parameters of a Gaussian mixture model) and adds a term to the cost function corresponding to the degree of uncertainty corresponding to a given trajectory in order to penalize control sequences with a high degree of uncertainty [17]. Finally, Risk-aware MPPI (RA-MPPI) samples disturbance realizations in addition to control actions and then approximately computes the Conditional Value-at-Risk (CVaR) for each control sequence using the

[1]Institute for Robotics and Intelligent Machines, Georgia Institute of Technology, Atlanta, GA 30332 USA. (email: jacobk@gatech.edu) This work was conducted during J. Knaup's internship at Honda Research Institute,USA.
[2]Honda Research Institute USA, Inc. (email: jovin_dsa@honda-ri.com, behdad_chalaki@honda-ri.com)

corresponding sampled trajectories and penalizes the control sequences that result in a high CVaR [18]. However, although there are several stochastic formulations of MPPI, there are no existing dual control formulations.

The contributions of this paper are twofold. First, we propose Dual Model Predictive Path-Integral Control (DMPPI), which extends MPPI by sampling model parameters and computing the expected cost over the *expected future* distribution of the parameters as opposed to the current parameter distribution. Second, we demonstrate the performance of our approach on a realistic highway on-ramp merging scenario featuring congested traffic using the high-fidelity vehicle simulation software IPG-CarMaker [19], thus, demonstrating the real-time computational capabilities of the algorithm in a multi-agent application. To the best of our knowledge, this is the first dual control formulation of MPPI and the first application of MPPI to autonomous highway merging.

II. Problem Formulation

We consider a general nonlinear stochastic system as

$$x_{t+1} = f(x_t, u_t, \bar{\theta}) + w_t, \tag{1}$$

where $x_t \in \mathbb{R}^{n_x}$ and $u_t \in \mathbb{R}^{n_u}$ denote the state and control action at time step $t \in \mathbb{N}$, respectively, while $\bar{\theta} \in \Theta \subseteq \mathbb{R}^{n_p}$ is a vector of constant but unknown parameters. Let $w_t \in \mathbb{R}^{n_x}$ be an i.i.d. random disturbance with corresponding density $a(w_t) = \mathcal{N}(0, \Sigma_w)$. System (1) can also be expressed using the lifted dynamics given by $x_k = g_{k-t}(x_t, u_{t:k-1}, \bar{\theta}, w_{t:k-1})$, for $k > t$ where $u_{t:k-1} = \{u_t, u_{t+1}, \dots, u_{k-1}\}$ and $w_{t:k-1} = \{w_t, w_{t+1}, \dots, w_{k-1}\}$. Given that the future state is contingent upon the realizations of the stochastic variable w_t, the dynamics can be expressed in terms of the conditional probability density, given by

$$x_k \sim h_{k-t}(x_k | x_t, u_{t:k-1}, \bar{\theta}). \tag{2}$$

Associated with (1), the cost function for the control horizon of length $N \in \mathbb{Z}^+$ is given by

$$J(x_{t+1:t+N}) = \phi(x_{t+N}) + \sum_{k=t+1}^{t+N-1} \ell_k(x_k), \tag{3}$$

where $x_{t+1:t+N} = \{x_{t+1}, x_{t+2}, \dots, x_{t+N}\}$, $\phi(\cdot)$ is the terminal cost, and $\ell_k(\cdot)$ the stage cost at time step k. Our goal is to derive an optimal control sequence $u_{t:t+N-1|t} = \{u_{t|t}, u_{t+1|t}, \dots, u_{t+N-1|t}\}$, where $u_{k|t}$ is the action planned for time step k at time step t, by solving the following optimization problem

$$\min_{u_{t:t+N-1|t}} \; S(x_t, u_{t:t+N-1|t}, \bar{\theta}), \tag{4}$$

where $S(\cdot)$ is defined as $S(x_t, u_{t:t+N-1|t}, \bar{\theta}) = \mathbb{E}_{w_{t:t+N-1} \sim a}[J(g_1(x_t, u_{t|t}, \bar{\theta}, w_t), \dots, g_N(x_t, u_{t:t+N-1|t}, \bar{\theta}, w_{t:t+N-1}))]$. Nonetheless, due to the unknown nature of $\bar{\theta}$, a direct solution to (4) is not feasible. Instead, we model $\bar{\theta}$ by $\theta \sim b(\theta)$, with $b(\cdot)$ denoting the probability distribution that represents our belief of $\bar{\theta}$. Rather than relying on a fixed belief distribution, we leverage Bayesian inference to estimate $\bar{\theta}$ online. Hence, the prior distribution is denoted as $b_0 = b(\theta)$, with the omission of the argument θ for brevity. The belief distribution is then updated based on the observed state as follows

$$b_{t+1}(\theta) = b(\theta | x_{0:t+1}, u_{0:t}) \propto b_t(\theta) h_1(x_{t+1} | x_t, u_t, \theta), \tag{5}$$

for $t = 0, 1, \dots$, where (5) is derived from Bayes' theorem (e.g., as in [20]).

The key question now concerns the approximation of (4) using (5). One straightforward method, denoted as Certainty Equivalence MPPI (CE-MPPI), is to, instead of (4), solve

$$\min_{u_{t:t+N-1|t}} \; S(x_t, u_{t:t+N-1|t}, \mathop{\mathbb{E}}_{\theta \sim b_t}[\theta]), \tag{6}$$

where we optimize with respect to the current expectation of the parameter using the previous data up to time step t. Another approach, similar to that in [16], we refer to as Ensemble MPPI (EMPPI), which solves

$$\min_{u_{t:t+N-1|t}} \; \mathop{\mathbb{E}}_{\theta \sim b_t}[S(x_t, u_{t:t+N-1|t}, \theta)], \tag{7}$$

through minimizing the expected cost over the current belief distribution of parameters at time t. However, both of the aforementioned methods fail to take into account the potential acquisition of future information within the planning horizon. In the event that additional information is acquired during the execution of the trajectory, resulting in the convergence of $b(\cdot)$ towards the true value $\bar{\theta}$, incorporating this new knowledge will result in a smaller deviation between the computed plan and the optimal solution of (4).

III. Proposed Approach

A. Dual Control Formulation

The proposed approach takes into consideration the information gain in the planning horizon by incorporating a Bayesian update step, thus establishing an implicit dual control framework. While the optimization problem (7) aims to minimize the expected cost based on the current belief state, our objective is to minimize the expected cost considering future belief states.

Therefore, we formulate the stochastic optimal control problem with the dual control framework as follows

$$\begin{aligned} &\min_{u_{t:t+N-1|t}} \; \bar{S}(x_t, u_{t:t+N-1|t}), \\ &\text{where } \bar{S}(x_t, u_{t:t+N-1|t}) \\ &\quad = \mathop{\mathbb{E}}_{\substack{\theta \sim b_{t+N}, \\ w_{t:t+N-1} \sim a}} [\phi(g_N(x_t, u_{t:t+N-1|t}, \theta, w_{t:t+N-1}))] \\ &\quad + \sum_{k=t+1}^{t+N-1} \mathop{\mathbb{E}}_{\substack{\theta \sim b_k, \\ w_{t:k-1} \sim a}} [\ell(g_{k-t}(x_t, u_{t:k-1|t}, \theta, w_{t:k-1}))], \end{aligned} \tag{8}$$

which computes the expected cost at each time step over the belief distribution using all information available at that time-step.

However, since the future belief distributions b_k for $k > t$ cannot be calculated ahead of time as the future state realizations are unknown, the control policy resulting from

Problem (8) is not causal. We instead use the approximate predicted future belief distributions $\hat{b}_{k|t}$, given by

$$\hat{b}_{k+1|t}(\theta) \propto \hat{b}_{k|t}(\theta) h_{k+1-t} \left(\mathop{\mathbb{E}}_{\substack{\tilde{\theta}\sim b_t, \\ w_{t:k}\sim a}} \left[g_{k+1-t}(x_t, u_{t:k}, \tilde{\theta}, w_{t:k}) \right] | x_t, u_{t:k}, \theta \right), \tag{9}$$

for all $k = t, \dots, t+N-1$ and where $\hat{b}_{t|t} = b_t$.

Thus, (9) yields the *causal* optimal control problem as

$$\begin{aligned} &\min_{u_{t:t+N-1|t}} \hat{S}(x_t, u_{t:t+N-1|t}), \\ &\text{where } \hat{S}(x_t, u_{t:t+N-1|t}) \\ &= \mathop{\mathbb{E}}_{\substack{\theta\sim\hat{b}_{t+N|t}, \\ w_{t:t+N-1}\sim a}} [\phi(g_N(x_t, u_{t:t+N-1|t}, \theta, w_{t:t+N-1}))] \\ &+ \sum_{k=t+1}^{t+N-1} \mathop{\mathbb{E}}_{\substack{\theta\sim\hat{b}_{k|t}, \\ w_{t:k-1}\sim a}} [\ell(g_{k-t}(x_t, u_{t:k-1|t}, \theta, w_{t:k-1}))], \end{aligned} \tag{10}$$

which computes the expected cost at each time step over the predicted future belief distributions.

Lemma 1. *The solution to Problem (10) is a causal control policy. That is, the optimal control applied at time t only depends on information available at or before time t.*

Proof. The proof follows trivially from (10) which computes the expected cost using the current state x_t and the planned future control sequence $u_{t:t+N-1|t}$ over the predicted future belief distributions. As seen in (9), the predicted belief dynamics are conditioned only on the current state x_t and planned control sequence $u_{t:t+N-1|t}$. This is in contrast to (8) which employs the true belief dynamics given by (5) that depend, however, on the future state realizations. □

B. Information-Theoretic MPPI

In practical applications, we address the computation of expectations in (9) and (10) through sampling-based approximations, as discussed further in Section III-C. The remaining challenge lies in solving Problem (10). Given that the optimization problem over the control actions is, in general, non-convex, we utilize MPPI [21], a sampling-based optimal control framework to solve (10).

Following the information-theoretic derivation of MPPI [21], we first define a distribution from which we can sample candidate control sequences $u_t \sim \mathcal{N}(\bar{u}_t, \Sigma_u)$, where $\mathcal{N}(\bar{u}_t, \Sigma_u)$ is a multivariate normal distribution with mean $\bar{u}_t \in \mathbb{R}^{n_u}$ and covariance $\Sigma_u \succ \mathbf{0}_{n_u \times n_u}$. Then, we may define the distribution $Q_{\bar{u}_{t:t+N-1|t}, \Sigma_u}$ for the control sequence $u_{t:t+N-1|t}$ with corresponding density given by $u_{t:t+N-1|t} \sim q(u_{t:t+N-1|t} | \bar{u}_{t:t+N-1|t}, \Sigma_u)$, where $\bar{u}_{t:t+N-1|t} = \{\bar{u}_{t|t}, \bar{u}_{t+1|t}, \dots, \bar{u}_{t+N-1|t}\}$ is the mean and Σ_u is the covariance.

Proposition 1 ([21]). *Control actions sampled from the optimal distribution Q^*, with corresponding density*

$$q^*(u_{t:t+N-1|t}) = \frac{1}{\eta} \exp\left(-\frac{\hat{S}(x_t, u_{t:t+N-1|t})}{\lambda}\right) p(u_{t:t+N-1|t}),$$

solve $\min_Q \mathbb{E}_Q[\hat{S}(x_t, u_{t:t+N-1|t})] + \lambda D_{\mathrm{KL}}(Q\|P)$, where λ denotes the inverse temperature, η is a normalizing constant, D_{KL} is Kullback–Leibler divergence, and P is an arbitrary base distribution with density $p(u_{t:t+N-1|t})$.

Thus, rather than directly optimizing (10), we seek to align the sample distribution as closely as possible with the optimal distribution by solving $u^*_{t:t+N-1|t} = \operatorname{argmin}_{\bar{u}_{t:t+N-1|t}} D_{\mathrm{KL}}(Q^* \| Q_{\bar{u}_{t:t+N-1|t}, \Sigma_u})$, which reduces to

$$u^*_{t:t+N-1|t} = \mathop{\mathbb{E}}_{Q_{\bar{u}_{t:t+N-1|t}, \Sigma_u}} [u_{t:t+N-1|t} \omega(u_{t:t+N-1|t})], \tag{11a}$$

$$\begin{aligned} \omega(u_{t:t+N-1|t}) = \frac{1}{\eta} \exp\Big(-\frac{1}{\lambda}\big(\hat{S}(x_t, u_{t:t+N-1|t}) \\ + \lambda \sum_{k=t}^{t+N-1} (u_{k|t} - \bar{u}_{k|t})^\top \Sigma_u^{-1} u_{k|t}\big)\Big). \end{aligned} \tag{11b}$$

C. Sampling-based Implementation

In order to approximate the expectations involving a, b, and q in (9), (10), (11), we employ importance sampling. We learn the belief distribution online using a particle filter. Initially, we generate N_p samples from $b(\theta)$ according to

$$\theta^i \sim b_0, \ i = 1, \dots, N_p, \tag{12}$$

and initialize the corresponding weights to $\nu^i_0 = 1/N_p$. The weights are then updated online based on the observed states using the following equations

$$\tilde{\nu}^i_{t+1} = \nu^i_t h_1(x_{t+1} | x_t, u_t, \theta^i), \tag{13a}$$

$$\nu^i_{t+1} = \tilde{\nu}^i_{t+1} / \sum_{i=1}^{N_p} \tilde{\nu}^i_{t+1}, \tag{13b}$$

where $h_1(x_{t+1}|x_t, u_t, \theta^i)$ is the conditional probability density for $x_{t+1} \sim \mathcal{N}(f(x_t, u_t, \theta^i), \Sigma_w)$ as in (2). This function represents the particle approximation of (5).

Remark 1. *Resampling may be added to (13) to enrich the sampled parameters. However, this is an implementation consideration rather than a theoretical one, and we found it unnecessary for the evaluation presented in Section IV.*

The predicted belief dynamics are approximated using the following predicted weights

$$\hat{\tilde{\nu}}^i_{k+1|t} = \hat{\nu}^i_{k|t} \hat{h}_{k+1-t}(\bar{x}_{k+1|t} | x_t, u_{t:k|t}, \theta^i), \tag{14a}$$

$$\hat{\nu}^i_{k+1|t} = \hat{\tilde{\nu}}^i_{k+1|t} / \sum_{i=1}^{N_p} \hat{\tilde{\nu}}^i_{k+1|t}. \tag{14b}$$

Here, $\hat{h}_{k+1-t}(\bar{x}_{k+1|t}|x_t, u_{t:k|t}, \theta^i)$ is an unbiased approximation of (2) given by the conditional probability density for $\bar{x}_{k+1|t} \sim \mathcal{N}(g_{k+1-t}(x_t, u_{t:k|t}, \theta^i), \Sigma^i_{x_{k+1|t}})$ such that

$$\bar{x}_{k+1|t} = \sum_{i=1}^{N_p} \nu^i_t \bar{x}^i_{k+1|t},$$

$$\Sigma^i_{x_{k+1|t}} = \sum_{j=1}^{N_w} (g_{k+1-t}(x_t, u_{t:k|t}, \theta^i, w^j_{t:k}) - \bar{x}^i_{k+1|t})(\star)^\top / N_w,$$

$$\bar{x}^i_{k+1|t} = \sum_{j=1}^{N_w} g_{k+1-t}(x_t, u_{t:k|t}, \theta^i, w^j_{t:k})/N_w,$$

where $(\star)$ represents repeated terms, $w^j_{t:k} = \{w^j_t, \dots, w^j_k\}$ and $w^j_k \sim a$, for $j = 1, \dots, N_w$. Thus, the objective function (10) may be approximated by

$$\begin{aligned}\hat{\hat{S}}&(x_t, u_{t:t+N-1|t}) \\ \approx \frac{1}{N_w}\sum_{i=1}^{N_p}\sum_{j=1}^{N_w}\Big[&\phi(g_N(x_t, u_{t:t+N-1|t}, \theta^i, w^j_{t:t+N-1}))\hat{\nu}^i_{k+N|t} \\ &+ \sum_{k=t+1}^{t+N-1} \ell(g_{k-t}(x_t, u_{t:k-1|t}, \theta^i, w^j_{t:k-1}))\hat{\nu}^i_{k|t}\Big]. \quad (15)\end{aligned}$$

Finally, we approximate the optimal control (11) using N_c control samples as follows

$$u^*_{t:t+N-1|t} = \sum_{\ell=1}^{N_c} u^\ell_{t:t+N-1|t}\hat{\omega}(u^\ell_{t:t+N-1|t}), \quad (16)$$

where $\hat{\omega}(u^\ell_{t:t+N-1|t}) = \frac{1}{\eta}\exp(-\frac{1}{\lambda}(\hat{\hat{S}}(x_t, u^\ell_{t:t+N-1|t}) + \lambda\sum_{k=t}^{t+N-1}(u^\ell_{k|t} - \bar{u}_{k|t})^\top\Sigma_u^{-1}u^\ell_{k|t}))$, $u^\ell_{t:t+N-1|t} \sim q(\cdot|\bar{u}_{t:t+N-1|t}, \Sigma_u)$ for all $\ell = 1, \dots, N_c$.

We summarize the proposed approach in Algorithm 1, which has a sampling complexity of $\mathcal{O}(N_c N_p N_w N)$.

Algorithm 1 Dual Model Predictive Path-Integral Control

Require: Sample sizes and control horizon: N_c, N_p, N_w, N
Require: MPPI distribution parameters: $\lambda, \bar{u}_{0:N-1|0}, \Sigma_u$
Require: Parameter prior distribution: $b_0 = b(\theta)$
$t \leftarrow 0$
$\theta^i \sim b_0$, $\nu^i_0 \leftarrow 1/N_p$, $\forall\, i = 1, \dots, N_p$
loop
  Sample $u^\ell_{t:t+N-1|t} \sim q(\cdot|\bar{u}_{t:t+N-1|t}, \Sigma_u)$, $\ell = 1, \dots, N_c$
  $w^j_{t:t+N-1} \sim a$, $j = 1, \dots, N_w$, $\hat{\nu}^i_{t|t} = \nu^i_t$, $i = 1, \dots, N_p$
  $\hat{\nu}^{i,\ell}_{k|t} \leftarrow$ Evaluate (14) using $x_t, u^\ell_{t:t+N-1|t}, \theta^i, w^j_{t:t+N-1}$, $\forall\, \ell = 1, \dots, N_c,\ i = 1, \dots, N_p,\ j = 1, \dots, N_w,\ k = t+1, \dots, t+N$
  $\hat{\hat{S}}^\ell \leftarrow$ Evaluate (15) using $x_t, u^\ell_{t:t+N-1|t}, \theta^i, \hat{\nu}^{i,\ell}_{k|t}, w^j_{t:t+N-1}$, $\forall\, \ell = 1, \dots, N_c,\ i = 1, \dots, N_p,\ j = 1, \dots, N_w,\ k = t, \dots, t+N$
  $u^*_{t:t+N-1|t} \leftarrow$ Evaluate (16) using $u^\ell_{t:t+N-1|t}, \hat{\hat{S}}^\ell, \ell = 1, \dots N_c$
  Apply $u_t = u^*_{t|t}$ and observe x_{t+1}
  $\nu^i_{t+1} \leftarrow$ Evaluate (13) using $\nu^i_t, x_{t+1}, x_t, u_t, \theta^i, i = 1, \dots, N_p$
  $t \leftarrow t+1$
end loop

Theorem 1. *The approximate solution to problem (10) computed from (16) preserves the dual control effect [13], that is, the planned control actions affect the entropy of the predicted future belief distribution.*

Proof. The proof follows from (14), in which the planned control sequence $u_{t:k|t}$ affects the weights $\hat{\nu}^i_{k+1|t}$ of the categorical distribution over $\{\theta^i\}_{i=1}^{N_p}$. □

IV. Autonomous Highway On-Ramp Merging Application

We evaluate the proposed approach (10) using the MPPI solution method (11) in a challenging highway on-ramp merge scenario featuring dense traffic conditions shown in Fig. 1. We use the DMPPI framework for interaction-aware *trajectory generation*, while low-level control is handled by the Vehicle Control System (VCS) of the Autonomous merging vehicle, also referred to as the ego vehicle. Therefore, the purpose of the DMPPI framework is primarily to design trajectories that will enable learning the traffic vehicles' behavior parameters and lead to a successful merge.

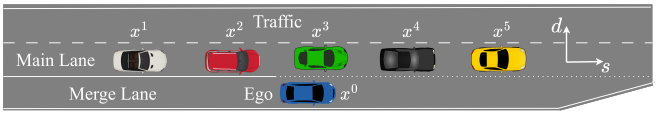


Fig. 1: Highway on-ramp merge scenario. Ego car is the autonomous merging car (in blue color) while all the main lane traffic cars are human driven vehicles.

We consider a highway on-ramp merging scenario with a single ego vehicle in the merge lane and n_v vehicles in the main road, where 1 and n_v denote the indices of the rear and lead vehicles, respectively. Let 0 be the index of the ego vehicle in the merge lane. At time step $k \in \mathbb{N}$, let $x^i_k = [v^i_{s,k}, v^i_{d,k}, s^i_k, d^i_k]^\top \in \mathbb{R}^4$ be the vector corresponding to the state of vehicle $i \in \{0, 1, \dots, n_v\}$, where $v^i_{s,k}$, $v^i_{d,k}$, s^i_k, and d^i_k denote longitudinal speed, lateral speed, longitudinal position, and lateral position, respectively in the Frenet coordinate frame. Let the stacked state vector of all vehicles at time step k be given by $x_k = [x^0_k, x^1_k, \cdots, x^{n_v}_k]^\top$. The control input of the ego vehicle is denoted by u_k, and we assume the state x_k is fully observable to the ego vehicle. We are interested in high-level, long-horizon motion planning, and assume that more complex low-level dynamics and control be handled by the autonomous vehicle's VCS. We therefore consider a double integrator dynamics, and let the control input of the ego vehicle be equal to its acceleration in the local Frenet coordinate frame, i.e., $u_k = [\hat{v}^0_{s,k}, \hat{v}^0_{d,k}]^\top \in \mathbb{R}^2$.

To predict how the vehicles on the main road behave, we employ the Merge Reactive Intelligent Driver Model (MR-IDM) [22]. This model extends the widely-used IDM [23] by also considering reaction to the merging vehicles in addition to the following vehicles in computing the longitudinal acceleration of the driver. We assume that main road vehicles do not make any lane changes, and thus, we only consider their longitudinal motion. Although we limit our scenario to highly-structured freeway driving, the proposed approach readily extends to more complex scenarios, provided a parameterized model for the behavior of other agents exists.

The combined ego-traffic dynamics are given by (1) with

$$f(x_t, u_t, \theta) = \mathcal{A}x_t + \mathcal{B}\begin{bmatrix}\sigma(x_t, u_t)\\ \zeta^1(x^1_t, x^0_t, x^2_t, \theta^1)\\ \vdots \\ \zeta^{n_v-1}(x^{n_v-1}_t, x^0_t, x^{n_v}_t, \theta^{n_v-1})\\ \zeta^{n_v}(x^{n_v}_t, x^0_t, \theta^{n_v})\end{bmatrix}, \quad (17)$$

where $\mathcal{A} = \text{blkdiag}(A, \dots, A)$, $\mathcal{B} = \text{blkdiag}(B, \dots, B)$,

$$\zeta^m = \begin{bmatrix}\dot{v}^m_s\\ \dot{v}^m_d\end{bmatrix} = \begin{bmatrix}\rho(x^m_t, x^0_t, x^{m+1}_t, \theta^m)\\ 0\end{bmatrix}, \quad (18a)$$

$$\zeta^{n_v} = \begin{bmatrix} \dot{v}_s^{n_v} \\ \dot{v}_d^{n_v} \end{bmatrix} = \begin{bmatrix} \tilde{\rho}(x_t^{n_v}, x_t^0, \theta^{n_v}) \\ 0 \end{bmatrix}, \tag{18b}$$

for $m \in \{1, 2, \ldots, n_v - 1\}$, and where $A \in \mathbb{R}^{4\times4}$ and $B^{4\times2}$ correspond to the standard second-order integrator dynamics. Let $\sigma(x_t, u_t) = [\dot{v}_{s,k}^0, \dot{v}_{d,k}^0]^\top$ in (17) be a clamping function which imposes kinematic constraints such as requiring $v_s^0 \geq 0$ and limiting acceleration magnitudes ($\dot{v}_{s_{\min}}^0 \leq \dot{v}_{s,k}^0 \leq \dot{v}_{s_{\max}}^0$, $\dot{v}_{d_{\min}}^0 \leq \dot{v}_{d,k}^0 \leq \dot{v}_{d_{\max}}^0$). Let ζ_t^i be the acceleration of traffic vehicle i for $i \in \{1, \ldots, n_v\}$, while $\rho(\cdot)$ and $\tilde{\rho}(\cdot)$ are the MR-IDM with and without a lead vehicle, respectively.[1] Note that ζ^{n_v} does not depend on a third vehicle, as x^{n_v} corresponds to the front-most traffic vehicle.

We set the cost function (3) at time step t for the control horizon of length $N \in \mathbb{Z}^+$ according to

$$\ell_k(x_k) = (x_k - x^g)^\top Q (x_k - x^g) + \ell^{\text{penalty}}(x_k), \tag{19a}$$
$$\phi(x_{t+N}) = (x_{t+N} - x^g)^\top Q_f(\star) + \ell^{\text{penalty}}(x_{t+N}), \tag{19b}$$
$$\ell^{\text{penalty}}(x_k) = q_I(\mathbf{1}^{\text{coll}}(x_k) + \mathbf{1}^{\text{road}}(x_k) + \mathbf{1}^{\text{inval}}(x_k)), \tag{19c}$$

for $k = \{t+1, t+2, \ldots, t+N-1\}$, where $x^g = [v^g, 0, 0, 0]^\top$ is the goal state, $Q = \text{diag}(q_{v_s}, q_{v_d}, q_s, q_d, 0, \ldots, 0) \in \mathbb{R}^{4(n_v+1)\times4(n_v+1)}$ is the state cost matrix, $Q_f = \text{diag}(0, 0, 0, q_d^f, 0, \ldots, 0) \in \mathbb{R}^{4(n_v+1)\times4(n_v+1)}$ is the terminal cost matrix, q_I is the violation penalty coefficient, and $\mathbf{1}^{\text{coll}}(x_k)$, $\mathbf{1}^{\text{road}}(x_k)$, $\mathbf{1}^{\text{inval}}(x_k)$ are the indicator functions for a collision with another vehicle, violating the road boundaries, or an improper merge (not between two vehicles), respectively.[2]

The selected parameters are shown in Table I. The cost function is designed to prioritize avoiding violations of the indicator functions, merging into the main lane of traffic, and maintaining the desired velocity. The terms are weighted so that if a merge cannot be completed safely, the lowest cost trajectory will be to avoid violating the penalty functions at the expense of neglecting to match the desired velocity or complete the merge.

TABLE I: DMPPI Parameters

Parameter	Value	Description
N_c	3,000	# of control samples
N_p^{pf}	10,000	# of parameter samples for particle filter
N_p^c	20	# of parameter samples for control
N_w	5	# of disturbance samples
N	50	# of control horizon time-steps
λ	10,000	Sharpness of MPPI optimal distribution
Σ_u	diag(10, 1.5)	Control sample covariance (m/s^2)^2
v^g	10	Target velocity m/s
q_{v_s}	10	Longitudinal velocity cost
q_{v_d}	0.1	Lateral velocity cost
q_s	0	Longitudinal position cost
q_d	10	Lateral position cost
q_d^f	10,000	Terminal lateral position cost
q_I	1,000,000	Violation penalty

[1]The equations for $\rho(\cdot)$ and $\tilde{\rho}(\cdot)$ are given in [22].

[2]As suggested in [21], we incorporate control constraints through clamping functions in the dynamics and task-related state constraints through weighted indicator penalty functions in the cost.

Remark 2. *An important implementation consideration is that, whereas the particle filter (12)-(13) and the predicted belief dynamics used for dual control (14)-(15) ideally utilize the same number of parameter samples N_p, in practice we use N_p^{pf} samples for the particle filter and N_p^c samples for control. We employ resampling to reduce the number of samples from the particle filter (which is relatively cheap) to DMPPI (which is more computationally expensive).*

We first evaluated the proposed approach in a low-fidelity highway on-ramp simulation environment that was developed in-house. We set up the evaluation to only include challenging merge scenarios requiring negotiation wherein the ego vehicle has less than 20 seconds and less than 300 meters from the soft-nose to the merge-ramp endpoint to successfully merge between five traffic vehicles, as shown in Fig. 1. The traffic vehicles are initialized with $v_{s_0}^m = 10$ m/s, $v_{d_0}^m = 0$ m/s, $d_0^m = 0$ m, and $s_0^{m+1} = s_0^m + 8$, where $s_0^1 = 0$, for $m = 1, \ldots, n_v$. That is, all traffic vehicles begin travelling at 10 m/s, 8 m apart from one another. The ego vehicle is initialized with $v_{s,0}^0 = 10$ m/s, $v_{d,0}^0 = 0$ m/s, $d_0^0 = -3.5$ m, and $s_0^0 \sim \mathcal{U}(s_0^1, s_0^{n_v})$, where $\mathcal{U}(s_0^1, s_0^{n_v})$ is a uniform distribution over the interval $[s_0^1, s_0^{n_v}]$. To robustly evaluate the proposed approach in comparison to the baselines, we randomly assign only one of the traffic vehicles to have a relatively friendly set of parameters of the MR-IDM, $\bar{\theta}^{\tilde{m}} \sim \Theta^{\text{friendly}}$, where $\tilde{m} \sim \{1, 2, \ldots, n_v\}$, thereby resulting in a scenario in which this specific vehicle will yield to the ego vehicle in case it attempts to merge in front of it. The rest of the vehicles have aggressive parameters, $\bar{\theta}^{\bar{m}} \sim \Theta^{\text{aggressive}}$, where $\bar{m} \neq \tilde{m}$, and will not yield. To further increase the difficulty of the scenario, merging behind the last car or ahead of the first car was treated as a failure case.

The proposed approach was compared against two state-of-the-art sampling-based baselines that do not feature active learning, but otherwise utilize the same particle filter and MPPI control solution approaches: CE-MPPI and EMPPI, where the objective $\hat{S}$ in (16) is replaced by the objectives of (6) and (7), respectively. All approaches use the same model, cost function, and parameters, where applicable. An important consideration is the design of the prior belief distribution over the parameters which we set to $b_0 = b(\theta) = 0.8\,\Theta^{\text{friendly}} + 0.2\,\Theta^{\text{aggressive}}$, where we optimistically bias the prior towards the friendly distribution to incentivize exploration. The merge success rate result for this low fidelity simulation test is shown in Table II.

TABLE II: Monte Carlo Merge Simulation Results

Method	CE-MPPI	EMPPI	**DMPPI**
Low-Fidelity Success Rate	0.44	0.47	0.70
High-Fidelity Success Rate	0.37	0.31	0.67
Avg. Min Long. Distance (m)	1.74	1.45	1.90
Avg. Min Lat. Distance (m)	1.07	1.06	1.14
Avg. Max Acceleration (m/s^2)	1.89	1.21	2.55

We next proceeded to evaluate the performance of the proposed approach in a real-time high-fidelity simulation environment using IPG-CarMaker [19] and ROS (Robot Operating System), as shown in Fig. 2. We set up the traffic

similar to the low-fidelity simulation setup to replicate the difficult merge scenarios. However, the highway used in this evaluation had a soft-nose to ramp-end distance of only 140 meters. To offset this added difficulty, we allowed up to two traffic vehicles to have parameters sampled from Θ^{friendly}. The DMPPI algorithm ran in real-time at approximately 10 Hz on a single NVIDIA RTX 3080 Ti GPU using Jax [24] for just-in-time compilation and hardware acceleration.

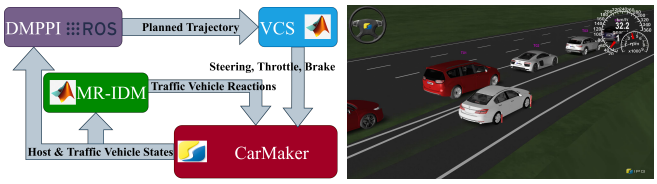
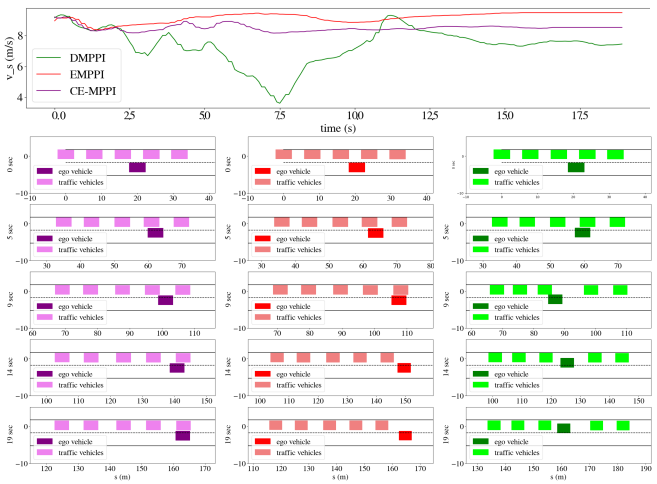


Fig. 2: IPG-CarMaker high-fidelity simulation environment.

The DMPPI control algorithm generates a planned trajectory using (17). Subsequently, this trajectory is broadcast via ROS to the VCS, which operates in the MATLAB-Simulink environment. The VCS computes the low-level control actions to attempt to follow the planned trajectory and sends them over ROS to CarMaker which then simulates the effects of these control actions on the ego vehicle. Additionally, the behavior of the traffic vehicles characterized by MR-IDM is processed within the MATLAB environment and the responses are broadcast to CarMaker over ROS, as shown in Fig. 2. The proposed approach was again compared against CE-MPPI and EMPPI. The rates of a successful merge (out of 100 trials) for each of the three approaches are shown in Table II, and a comparison of a single trial may be seen in Fig. 3.



(a) CE-MPPI (b) EMPPI (c) DMPPI

Fig. 3: Comparison of snapshots of trajectories and velocity profiles from a single Monte Carlo merge scenario trial.

Although DMPPI has a higher success rate over the baselines, it does not compromise safety as all three methods attained a 0% collision rate and maintained similar distances from the traffic vehicles as seen in Table II, where the 4$^{\text{th}}$ and 5$^{\text{th}}$ rows refer to the minimum longitudinal and lateral distances between the ego vehicle and any traffic vehicle averaged over all Monte Carlo trials. Rather, the superior performance is attained by intelligently expending more control effort in order to probe the system as seen in the last row of Table II, which refers to the average over the Monte Carlo trials of the maximum absolute acceleration of the ego vehicle for each trial. This resulting higher acceleration for DMPPI is still lower than the common acceleration limits for AD/ADAS systems [25]. Moreover, the 0% collision rate as seen from our simulation study demonstrates that the soft constraints imposed through the penalty functions are effectively enforced.

The relative performance between the three approaches is comparable to that for the low-fidelity simulation. However, the overall performance for all approaches is slightly reduced in the high-fidelity simulation, which is due, in part, to the model mismatch between the environment and the model used for control design and parameter fitting as well as due to the additional kinematic and dynamic limitations imposed by the high-fidelity vehicle model.

The explanation for the superior merge success rate of DMPPI may be seen through Fig. 3, which shows snapshots at different time instants for a single trial. Both CE-MPPI and EMPPI algorithms tend to maintain a relatively constant velocity since no merging gap is available and these observations fail to provide any indication of a forthcoming gap opening up for merging. DMPPI, on the other hand, actively probes the vehicles, by adjusting its velocity, in order to better identify the driver's behavior and find a friendly driver in front of whom to merge. It may be seen that, initially, all three approaches follow the same trajectory. However, once it is seen that a driver will not yield, in this case, DMPPI sharply slows down to probe another driver, while the other two approaches maintain their velocity and relative positions to the traffic vehicles waiting for a feasible gap to open. By decelerating, DMPPI shifts its position to better identify the parameters of another driver, hoping that they will be friendly and willing to yield. By doing so, a yielding vehicle is found and the merge may be completed.

V. Conclusion

In this paper, we introduced the novel Dual Model Predictive Path-Integral (DMPPI) control algorithm, a sampling-based implicit dual optimal control framework. We employ this framework in a highway on-ramp merge scenario in which our algorithm is used to actively learn the unknown behavior of other drivers in order to successfully merge into congested freeway traffic. We evaluated the performance of our framework in a high-fidelity simulation environment with the algorithm running in real-time at 10 Hz. The proposed approach demonstrated superior performance over two state-of-the-art variations of MPPI that rely on passive learning. Future work may include more rigorous integration between MPPI and the VCS, evaluation on more complex on-ramps and traffic scenarios generated from real-life conditions, and deployment on a physical autonomous vehicle.

REFERENCES

[1] L. Claussmann, M. Revilloud, D. Gruyer, and S. Glaser, "A review of motion planning for highway autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 5, pp. 1826–1848, 2019.

[2] R. Song and B. Li, "Surrounding vehicles' lane change maneuver prediction and detection for intelligent vehicles: A comprehensive review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 6046–6062, 2021.

[3] J. Rios-Torres and A. A. Malikopoulos, "A survey on the coordination of connected and automated vehicles at intersections and merging at highway on-ramps," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 5, pp. 1066–1077, 2016.

[4] A. Zgonnikov, D. Abbink, and G. Markkula, "Should I stay or should I go? Evidence accumulation drives decision making in human drivers," *PsyArXiv*, 2020, doi: 10.31234/osf.io/p8dxn.

[5] S. A. Fernandez, M. A. Marinho, M. Vakilzadeh, and A. Vinel, "Highway on-ramp merging for mixed traffic: Recent advances and future trends," in *IEEE 29th International Conference on Network Protocols (ICNP)*, Dallas, TX, Nov. 1–5, 2021, pp. 1–6.

[6] H. Wang, S. Yuan, M. Guo, X. Li, and W. Lan, "A deep reinforcement learning-based approach for autonomous driving in highway on-ramp merge," *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, vol. 235, no. 10-11, pp. 2726–2739, 2021.

[7] M. Schuurmans, A. Katriniok, C. Meissen, H. E. Tseng, and P. Patrinos, "Safe, learning-based MPC for highway driving under lane-change uncertainty: A distributionally robust approach," *Artificial Intelligence*, vol. 320, p. 103920, 2023.

[8] W. Liu, Y. Zhai, G. Chen, and A. Knoll, "Gaussian process based model predictive control for overtaking scenarios at highway curves," in *IEEE Intelligent Vehicles Symposium (IV)*, Aachen, Germany, June 5–9, 2022, pp. 1161–1167.

[9] B. Evens, M. Schuurmans, and P. Patrinos, "Learning MPC for interaction-aware autonomous driving: A game-theoretic approach," in *European Control Conference (ECC)*, London, UK, July 12–15, 2022, pp. 34–39.

[10] H. Kimura, M. Takahashi, K. Nishiwaki, and M. Iezawa, "Decision-making based on reinforcement learning and model predictive control considering space generation for highway on-ramp merging," *IFAC-PapersOnLine*, vol. 55, no. 27, pp. 241–246, 2022.

[11] N. Albarella, D. G. Lui, A. Petrillo, and S. Santini, "A hybrid deep reinforcement learning and optimal control architecture for autonomous highway driving," *Energies*, vol. 16, no. 8, p. 3490, 2023.

[12] J. Lubars, H. Gupta, S. Chinchali, L. Li, A. Raja, R. Srikant, and X. Wu, "Combining reinforcement learning with model predictive control for on-ramp merging," in *IEEE International Intelligent Transportation Systems Conference (ITSC)*, Indianapolis, IN, Sept 19–22, 2021, pp. 942–947.

[13] H. Hu and J. F. Fisac, "Active uncertainty reduction for human-robot interaction: An implicit dual control approach," in *International Workshop on the Algorithmic Foundations of Robotics*. College Park, MD: Springer, June 22–24, 2022, pp. 385–401.

[14] S. H. Nair, V. Govindarajan, T. Lin, Y. Wang, E. H. Tseng, and F. Borrelli, "Stochastic MPC with dual control for autonomous driving with multi-modal interaction-aware predictions," *arXiv preprint arXiv:2208.03525*, 2022.

[15] G. Williams, N. Wagener, B. Goldfain, P. Drews, J. M. Rehg, B. Boots, and E. A. Theodorou, "Information theoretic MPC for model-based reinforcement learning," in *IEEE International Conference on Robotics and Automation (ICRA)*, Singapore, May 29–June 3, 2017, pp. 1714–1721.

[16] I. Abraham, A. Handa, N. Ratliff, K. Lowrey, T. D. Murphey, and D. Fox, "Model-based generalization under parameter uncertainty using path integral control," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2864–2871, 2020.

[17] E. Arruda, M. J. Mathew, M. Kopicki, M. Mistry, M. Azad, and J. L. Wyatt, "Uncertainty averse pushing with model predictive path integral control," in *IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*, Birmingham, UK, Nov. 15–17, 2017, pp. 497–502.

[18] J. Yin, Z. Zhang, and P. Tsiotras, "Risk-aware model predictive path integral control using conditional value-at-risk," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, London, UK, May 29–June 2, 2023, pp. 7937–7943.

[19] C. IPG, "Reference manual version 9.0. 1," *IPG Automotive GmbH: Karlsruhe, Germany*, 2021.

[20] M. Speekenbrink, "A tutorial on particle filters," *Journal of Mathematical Psychology*, vol. 73, pp. 140–152, 2016.

[21] G. Williams, P. Drews, B. Goldfain, J. M. Rehg, and E. A. Theodorou, "Information-theoretic model predictive control: Theory and applications to autonomous driving," *IEEE Transactions on Robotics*, vol. 34, no. 6, pp. 1603–1622, 2018.

[22] D. Holley, J. D'sa, H. N. Mahjoub, G. Ali, B. Chalaki, and E. Moradi-Pari, "MR-IDM–merge reactive intelligent driver model: Towards enhancing laterally aware car-following models," *arXiv preprint arXiv:2305.12014*, 2023.

[23] M. Treiber, A. Hennecke, and D. Helbing, "Congested traffic states in empirical observations and microscopic simulations," *Physical review E*, vol. 62, no. 2, p. 1805, 2000.

[24] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang, "JAX: composable transformations of Python+NumPy programs," 2018. [Online]. Available: http://github.com/google/jax

[25] K. N. de Winkel, T. Irmak, R. Happee, and B. Shyrokau, "Standards for passenger comfort in automated vehicles: Acceleration and jerk," *Applied Ergonomics*, vol. 106, p. 103881, 2023.