Uncertainty-aware Contact-safe Model-based Reinforcement Learning

Cheng-Yu Kuo1*, Andreas Schaarschmidt2, Yunduan Cui3, Tamim Asfour2, and Takamitsu Matsubara1

Abstract—This paper presents contact-safe Model-based Reinforcement Learning (MBRL) for robot applications that achieves contact-safe behaviors in the learning process. In typical MBRL, we cannot expect the data-driven model to generate accurate and reliable policies to the intended robotic tasks during the learning process due to data scarcity. Operating these unreliable policies in a contact-rich environment could cause damage to the robot and its surroundings. To alleviate the risk of causing damage through unexpected intensive physical contacts, we present the contact-safe MBRL that associates the probabilistic Model Predictive Control’s (pMPC) control limits with the model uncertainty so that the allowed acceleration of controlled behavior is adjusted according to learning progress. Control planning with such uncertainty-aware control limits is formulated as a deterministic MPC problem using a computationally-efficient approximated GP dynamics and an approximated inference technique. Our approach’s effectiveness is evaluated through bowl mixing tasks with simulated and real robots, scooping tasks with a real robot as examples of contact-rich manipulation skills. (video: https://youtu.be/8uTDYYUKeFM)

I. INTRODUCTION

Model-based Reinforcement Learning (MBRL) [1] [2] is attractive in robotics scenarios due to its effectiveness and sample-efficiency. However, we cannot expect the data-driven model to generate reliable control to the intended robotic task during the learning process while sample scarcity. Although applying these unreliable controls are required as exploration, it could damage the robot and its surroundings in a contact-rich environment, for instance, the kitchen. Since safety is the primary consideration, a contact-safe learning process in MBRL needs to be considered.

Several prior methods have addressed accomplishing such contact-rich manipulation tasks with Reinforcement Learning (RL): learn a torque profile to perform picking up objects manipulation tasks with imitation learning and refines with RL [4]. A recurrent neural network model is employed to perform food cutting tasks with MBRL [5]. Although achieving marvelous results, contacts’ safety during the learning process was not considered.

Contact-rich tasks include contacts’ behaviors in which accurate prediction is hard, and its long-term prediction, needed in MBRL, is harder [6]. To alleviate this difficulty, Model Predictive Control (MPC) [7] [8] may be a reasonable choice. The MPC truncates the prediction length short, and recursively solve the problems with the newly observed state. It is essentially robust for modeling errors. Nevertheless, it may not be sufficient for acquiring contact-rich manipulations due to the lack of a contact-safe exploration mechanism.

Tactile exploration by touch for an unknown object exemplifies the contact-safe behavior. Without sufficient knowledge of the object, human beings’ fundamental fear or intolerance to the uncertainty causes cautious behaviors while approaching the object for maximizing survival [9] [10]. Inspired by this concept, associating the uncertainties to limit the robot’s control signal can be a reasonable approach to contact-safe exploration. A similar method adjusts the robot’s stiffness with a risk indicator while interacting with human operators [11], where the operator’s head position and orientation determine the risk.

This paper presents the contact-safe MBRL that can alleviate the risk of unexpected and damaging physical contacts during the learning process. As illustrated in Fig. 1 the control limits in the probabilistic-MPC (pMPC) problem are automatically adjusted to the learning progress by an uncertainty-aware approach that associates the control limits with the model-uncertainty, as the state’s predictive distribution. Prior research had addressed to handle constraint in pMPC [12]; however, not in a principle way where numerical issues may occur during control planning [13] [7]. Therefore, implementation of control planning with such uncertainty-aware control limits is formulated as a deterministic MPC problem [14] using a computationally efficient pMPC [15] that support dynamics control limits by exploiting Pontryagin’s Maximum Principle (PMP) [16].

The contributions of this work are the followings:

(1) Present the contact-safe MBRL that achieves contact-safe
behaviors by utilizing the model-uncertainty in handling the control limits for pMPC.

(2) Empirically demonstrate the contact-safe behaviors through bowl mixing tasks with a simulated and real robot, and scooping tasks with a real robot as examples of contact-rich manipulation skills.

This paper has the following structure: Section II provides insights in related works. This is followed by basic knowledge of this work in Section III. The Section IV details the contact-safe MBRL. With results in Section V the effectiveness is verified. The paper is concluded with discussions, Section VI.

II. RELATED WORK

A. Contact Safety in Contact-rich Manipulations with RL

Unlike contact-free manipulations, where most of the tasks could be addressed with position-controlled rigid robots [17], there are two common approaches to avoid intensive collisions in contact-rich scenarios. The first approach consists of learning a controller or torque profile to adjust robots’ stiffness with RL. For instance, 1) performing a variety of industrial insertion tasks with visual inputs [18]. 2) Learn the controller to perform a force-sensitive peg-in-hole task in sub-millimeter tolerances [19]. 3) Performing fail-safe ring-insertion and peg-insertion tasks by learning the controller gain of the parallel position/force control [20]. 4) Perform a peg-in-hole task with an RL learned controller to switch to a RL policy that minimize the uncertainty [12] [26]. Lee et al. [27] utilized such computation-efficient nature by modeling the system dynamics with LGM-FF, an approximated GPs model via Linear Gaussian Model [28] with the Fastfood random features [29]. The Fastfood generates the frequency components of a Fourier transformed kernel expression by following the Fubini’s theorem [30].

State predictions with a probabilistic system dynamics are considerably more complicated than a deterministic approach. Following the Fubini’s Theorem, an analytic moment-matching method approximates future state as a Gaussian distribution by propagating uncertainty [12].

The computation cost of modeling the system dynamics with LGM-FF is \(O(M^3)\), and \(O(M^2)\) in exploiting moment-matching, where \(M\) is the number of Fastfood features. In contrast, modeling the system dynamics and exploit moment-matching with standard GPs is \(O(N^3)\) and \(O(N^2)\). The efficiency of Fastfood feature selection is proved [29], and demonstrated that \(M \ll N\) in robotic task acquisitions [15]. We utilized such computation-efficient nature by modeling the system dynamics with LGM-FF.

B. Utilizing Uncertainties in Probabilistic Approaches

The uncertainty in a probabilistic approach provides rich information, several prior works had utilized such information for various purposes. Some of them utilized the uncertainty in the way that encourages the agent to expand its exploration coverage [24] [25]. In contrast, some others plan for a policy that minimizes the uncertainty [12] [26]. Lee et al. [27] proposed a guided uncertainty-aware approach for guiding the robot to an uncertainty area and switch to a RL policy to perform a peg-in-hole task in such area.

Apart from these applications, our method explores a novel application of model-uncertainty from above existing studies in a sense that the model-uncertainty is associated with control limits in adjusting agent’s exploration behaviors to achieve the aimed contact-safe behaviors.

III. PRELIMINARY

A. Overview

Several pieces of research had demonstrated the sample-efficiency of MBRL with Gaussian Processes (GPs) dynamics [12] [14]. However, despite the excellent data-efficiency, modeling standard GPs dynamics largely suffers from computation inefficiency while sample size \(N\) increases through the learning process. Resulting in a trade-off between control frequency and maximum allowed sample size, which may not be suitable for real robot applications since control frequency and samples for modeling dynamics are equally important. A sample-and-computational-efficient approach, the SCP-MPC [15], is proposed to alleviate this trade-off by modeling the dynamics with LGM-FF, an approximated GPs model via Linear Gaussian Model [28] with the Fastfood random features [29]. The Fastfood generates the frequency components of a Fourier transformed kernel expression by following the Bochner’s theorem [30].

The computation cost of modeling the system dynamics with LGM-FF is \(O(M^3)\), and \(O(M^2)\) in exploiting moment-matching, where \(M\) is the number of Fastfood features. In contrast, modeling the system dynamics and exploit moment-matching with standard GPs is \(O(N^3)\) and \(O(N^2)\). The efficiency of Fastfood feature selection is proved [29], and demonstrated that \(M \ll N\) in robotic task acquisitions [15]. We utilized such computation-efficient nature by modeling the system dynamics with LGM-FF.

B. LGM-FF Dynamics

Given a state input \(x_t \in \mathbb{X} \subset \mathbb{R}^D\) and a control signal \(u_t \in \mathbb{U} \subset \mathbb{R}^U\), consider a GPs dynamics \(f: \mathbb{X} \times \mathbb{U} \rightarrow \mathbb{X}\) representing the underlying true dynamics of the system:

\[
x_{t+1} = f(x_t, u_t) + \epsilon,
\]

where \(\epsilon \sim N(0, \Sigma_n)\) is the system noise with covariance matrix \(\Sigma_n = \text{diag}([\sigma_{n,1}^2, ..., \sigma_{n,D}^2])\).

For each target dimension \(a = 1, ..., D\), we approximately represent each GPs dynamics \(f_a(\cdot)\) via LGM [28] with corresponding weight \(w_a \in \mathbb{R}^{M \times D}\). The LGM model is then trained with \(N\) samples collected from the system: \(\mathcal{X} = [\tilde{x}_1, ..., \tilde{x}_N]\) and \(Y = [y_1, ..., y_N]\) that contain individual training input tuple \(\tilde{x}_t := [x_t, u_t]^\top\) and target \(y_t := x_{t+1}\). We obtain the predictive distribution of a new input \(\hat{x}\) as:

\[
p(f_a(\hat{x})|\mathcal{X}, \mathcal{Y}) \sim N(\hat{\Phi}^\top \Phi^\star \Phi^\star \top A^{-1} \Phi^\star, \Phi^\star \top A^{-1} \Phi^\star),
\]

where \(\Phi^\star : \mathbb{X} \times \mathbb{U} \rightarrow \mathbb{R}^M\) is the Fastfood feature map and \(\Phi_a^\star := \Phi(x_a)\). By applying maximum a posteriori and maximum marginal likelihood estimation [31] [32], we obtain
optimal parameters as:
\[
\begin{align*}
\mathbf{w}_a &= \sigma_{n,a}^{-2} \mathbf{A}_a^{-1} \Phi_a \left( \tilde{\mathbf{X}} \right) \mathbf{y}_a; \\
\mathbf{A}_a &= \sigma_{n,a}^{-2} \Phi_a \left( \tilde{\mathbf{X}} \right) \Phi_a \left( \tilde{\mathbf{X}} \right)^\top + \sigma_{n,a}^{-2} \mathbf{I},
\end{align*}
\]
where \(\sigma_{n,a}^{-2}\) is the signal noise, and \(\mathbf{I}\) is the identity matrix.

\[\text{C. Moment-matching Prediction with LGM-FF Given Uncertain Inputs}\]

We assume each latent prediction is independent with zero covariances. Therefore, the approximated probabilistic state \(\mathbf{p} (\mathbf{x}_t) \approx N (\mathbf{\mu}_t, \mathbf{\Sigma}_t)\) has the following form: \(\mathbf{\mu} = [\mu_1, ..., \mu_D]^\top\), and \(\mathbf{\Sigma} = \text{diag} [\sigma_1^2, ..., \sigma_D^2]\).

Given a probabilistic state \(\mathbf{x}_t\) and a deterministic control \(\mathbf{u}_t\), we obtain the approximated predictive distribution \(\mathbf{p} (\mathbf{x}_{t+1}) \approx N (\mathbf{\mu}_{t+1}, \mathbf{\Sigma}_{t+1})\) by integrating the LGM-FF dynamics throughout the inputted state distribution \(\mathbf{p} (\mathbf{x}_t)\). Assuming independent operations between latent models, the indicator \(a\) is omitted to simplify notations. For each latent model, we obtain the predictive \(\mathbf{\mu}_{t+1}\) and \(\sigma_{t+1}^2\) by
\[
\begin{align*}
\mathbf{\mu}_{t+1} &= \mathbb{E}_{\tilde{\mathbf{x}}_t} [ f (\tilde{\mathbf{x}}_t) | \mathbf{\mu}_t, \mathbf{\Sigma}_t ] = \mathbf{w} \mathbf{q}; \\
\sigma_{t+1}^2 &= \text{var} [ f (\tilde{\mathbf{x}}_t) | \mathbf{\mu}_t, \mathbf{\Sigma}_t ] \\
&= \text{tr} (\mathbf{A}^{-1} \mathbf{Q}) + \mathbf{w} \mathbf{Q} \mathbf{w} - \mu_{t+1}^2,
\end{align*}
\]
with analytic solutions of \(\mathbf{q} = \int \Phi (\tilde{\mathbf{x}}_t) \mathbf{p} (\mathbf{x}_t) \mathbf{dx}_t\) and \(\mathbf{Q} = \int \Phi (\tilde{\mathbf{x}}_t) \Phi (\tilde{\mathbf{x}}_t)^\top \mathbf{p} (\mathbf{x}_t) \mathbf{dx}_t\) are provided in [15].

\[\text{IV. CONTACT-SAFE MBRL}\]

\[\text{A. Overview}\]

The contact-safe MBRL’s learning process contains \(N_{\text{trial}}\) trials with \(L_{\text{rollout}}\) rollouts (steps) each to perform task acquisitions. At each rollout, the pMPC plans for an open-loop \(H\) step optimal control sequence \(\mathbf{u}_1, ..., \mathbf{u}_H\) by recursively exploit moment-matching predictions and apply the optimal control to the environment with corresponding sample collecting. Before each trial ends, the LGM dynamics are updated with all previously collected samples.

During the pMPC control planning, the predictive state distribution of a data-driven probabilistic dynamics model will return to its prior belief while scarce samples, or converge to observations while samples are adequate. Therefore, we can view the model-uncertainty, each predictive distribution, to measure how confident the agent knows about its dynamics under the current standpoint.

Given previous knowledge about the model-uncertainty and the pMPC, we cannot expect the pMPC to plan a reliable control signal under high model-uncertainty, where the predictions are imprecise due to the sample scarcity. However, applying these control signals to the environment and observe consequences are necessary for the agent to acquire its dynamics in MBRL. Exquisitely, we anticipate that the agent can explore cautiously to prevent intensive contacts, and exploit confidently to maximize performance.

With these in mind, we present the contact-safe MBRL, an MBRL algorithm that adaptively adjusts agent’s awareness by associating model-uncertainty in handling pMPC control limits. The contact-safe MBRL promotes two safe-characteristics:

(A) The agent takes small/gentle actions while high uncertainty to explore carefully.

(B) Encourage the agent to decrease its momentum while the uncertainty increased.

The characteristic (A) is straightforward that small/gentle actions are generally “safer” than big/intensive actions during the exploration. The characteristic (B) is needed because small actions also indicate small changes in momentum, causing the agent might be incapable of reducing its momentum effectively during high uncertainty.

There are two benefits of achieving the objective with an uncertainty-aware approach: 1) The model-uncertainty decreases through trials while samples are becoming adequate. An uncertainty-aware approach can seamlessly reduce the agent’s awareness for better performance. 2) The model-uncertainty varies from individual predictions. Thus, the agent can react to sudden changes in uncertainty and adjust its awareness attentively.

\[\text{B. The Contact-safe MBRL Framework}\]

\[\text{1) Deterministic Reformulated System Dynamics:}\]

For associating the state’s uncertainty in handling pMPC control limits by exploiting PMP [16], we reformulate the LGM-FF into a deterministic system dynamics representation \(f_d (\cdot)\) [14]. Predictions with the deterministic dynamics \(f_d (\cdot)\) follow the moment-matching uncertainty propagation, Eq. (5), with a expanded formulation of state \(z_t\) that accommodates both mean \(\mu_t\) and variance \(\Sigma_t\) as its entries:
\[
z_{t+1} = f_d (\tilde{z}_t) \in \mathbb{R}^{2D+U} \rightarrow \mathbb{R}^D
\]
where \(\tilde{z}_t := [z_t^\top, \mathbf{u}_t^\top]^\top \in \mathbb{R}^{2D+U}\) with \(z_t := [\mu_t^\top, 1]^\top \in \mathbb{Z}\), and \(\mathbf{1}\) denotes a column vector of ones.

For the following robotic applications, we define the deterministic reformulated state in Eq. (7) as:
\[
\mu = [\mu_p^\top, \mu_p^\top]^\top, \Sigma = \text{diag} [\Sigma_p, \Sigma_p],
\]
where \(\mathbf{p}\) and \(\mathbf{p}\) are the position and velocity of each actuator, with its mean and covariance matrices \(\mu_p, \mu_p\) and \(\Sigma_p, \Sigma_p\).

\[\text{2) pMPC with Uncertainty-aware Control Limits}\]

Associating the predictive state \(z\) that accommodates variance, we handle the pMPC control limits with the uncertainty for all \(H\) step via functions \(u_{\text{min}} (z_k), u_{\text{max}} (z_k)\):
\[
\begin{align*}
\text{minimize} & \quad \mathcal{L}_d = \sum_{k=2}^{H+1} \ell_d (z_k) \\
\text{subject to} & \quad z_{k+1} = f_d (z_k) \\
& \quad z_k = [z_k^\top, u_k^\top]^\top, \quad z_1 = z_t \\
& \quad z_k \in \mathbb{Z}, \quad k = 1, 2, ..., H + 1 \\
& \quad u_k \in [u_{\text{min}} (z_k), u_{\text{max}} (z_k)], \\
& \quad k = 1, 2, ..., H
\end{align*}
\]
where \(\ell_d : \mathbb{Z} \times \mathbb{U} \rightarrow \mathbb{R}\) is the deterministic immediate loss.
3) Control Limits Function with Tunable Awareness:
The objective of \( u_{min}(z_k), u_{max}(z_k) : \mathbb{Z} \to \mathbb{U} \) in Eq. (5) are handling the control limits such that the system can meet the two safe-characteristics. Therefore, we propose a control limits scaling and translating mechanism:

\[
\begin{align*}
  u_{max}(z_k) &= u_{max}K_s(z_k)I - K_i(z_k)I \\
  u_{min}(z_k) &= u_{min}K_s(z_k)I - K_i(z_k)I
\end{align*}
\] (10) (11)

where \( K_s(\cdot) : \mathbb{Z} \to \mathbb{R} \) associate the state \( z_k \) to scale and translate the feasible control boundary of the agent \( \mathbb{U} \in [u_{min}, u_{max}] \). Both \( u_{min} \) and \( u_{max} \) are identically adjusted.

The scaling contributes to the safe-characteristic (A) by scaling down the control limits during high uncertainty. To attain this, we are seeking for a mapping that \( K_s(z_k) \)'s value:

- (A1) decreases while the uncertainty increases;
- (A2) has a pre-defined minimum value within \((0, 1]\) to prevent control limits with zero or negative range; and
- (A3) has a maximum value of 1.0 during low uncertainty to return to its feasible boundary \([u_{min}, u_{max}]\);
- (A4) easy tunability to adjust the uncertainty-awareness.

The translating contributes to the safe-characteristic (B) by translating the control limits to the negative direction of the agent's velocity during high uncertainties. To attain this, we are seeking for a mapping that \( K_t(z_k) \)'s value:

- (B1) increase with velocity only if the uncertainty increases;
- (B2) has a minimum value of 0 during low uncertainty to return to its feasible boundary \([u_{min}, u_{max}]\); and
- (B3) easy tunability to adjust the uncertainty-awareness and velocity sensitivity.

With the above in mind, and the ease of creating increasing or decreasing mappings within the range of \([0, 1]\) under \( \mathbb{R}_{\geq 0} \) inputs, we designed the mapping with exponential curves:

\[
\begin{align*}
  K_s(z) &= (1 - \beta_s) \exp(-\alpha_s \|\Sigma_p\|_2) + \beta_s, \quad (A1) \\
  &= \begin{cases} 
    \exp(-\alpha_s \|\Sigma_p\|_2) & (A3) \\
    \exp(-\alpha_s \|\Sigma_p\|_2) & (A4)
  \end{cases} \quad (A2) \\
  K_t(z) &= \begin{cases} 
    \exp(-\alpha_t \|\Sigma_p\|_2) \gamma_t \mu_p & (B1) \\
    \exp(-\alpha_t \|\Sigma_p\|_2) \gamma_t \mu_p & (B2)
  \end{cases} \quad (B3)
\end{align*}
\] (12) (13)

for a consistent scale of uncertainty, only the position variance \( \Sigma_p \) is measured to estimate the uncertainty at the current state. The \( \alpha_s, \alpha_t \in \mathbb{R}_{\geq 0} \) are the tunable parameter for uncertainty-awareness; the \( \gamma_t \in \mathbb{R}_{>0} \) is the sensitivity to the velocity, which usually set as the ratio of the control's feasible upper bound to the agent's max velocity to prevent over translating; the \( \beta_s \) is the minimum value of scaling. Fig[2] illustrates the scaling and translating mechanism.

By tuning the uncertainty-awareness \( \alpha_s \) and \( \alpha_t \), agent’s exploration cautiousness is adjusted. For instance, contact-safe control limits disabled with \( \alpha_s, \alpha_t = 0 \), or enabled with excessive uncertainty-awareness \( \ln \alpha_s, \ln \alpha_t = 10 \). Fig. [3] exemplifies the relationship between the state \( z \) and the \( K_s(\cdot), K_t(\cdot) \), with \( \beta_s, \gamma_t = 0.2 \). Note that the translating is active only if \( \mu_p \neq 0 \) under certain level of uncertainty.

4) Ahead Control Planning for Delay Compensation:
The MBRL framework we applied in our previous work, SCP-MPC [15], suffers from the inconsistent active control signal between states of each time step. The inconsistency further impacts on the accuracy of the LGM dynamics. To alleviate the inconsistency, similar to the parallel MPC workflow [33], we introduce a one-step-ahead pMPC control planning. At each time step \( t \), we exploit an one-step state prediction \( \tilde{z}_{t+1} = f_{\tilde{d}}(\tilde{z}_t) \) and the pMPC plans for the optimal one-step-ahead control \( u_{t+1} \) based on the predicted state \( \tilde{z}_{t+1} \). The ahead predictions compensates the delay between observing \( z_t \) and operating \( u_t \), resulting a consistent control between \( z_t \) and \( z_{t+1} \). Fig. [4] There are three benefits of applying the ahead prediction. First, increased consistency between states. Second, the control frequency is increased since the pausing after operating control is no longer needed. Third, the variances of the predicted state \( \tilde{z}_{t+1} \) estimates the uncertainty of the following state \( z_{t+1} \), which can be utilized in control limits handling.

5) Summary:
Algorithm [1] summarizes the contact-safe MBRL. First, the deterministic reformulated system dynamics allow us to associate the pMPC control limits with the state’s uncertainty in principle way. Next, we designed the control limits function with a tunable uncertainty-awareness to meet the two safe-characteristics. Finally, ahead control planning is added to the framework to compensate for the control delay.
Scooping task with real 4-DoF arm.

Algorithm 1: Contact-safe MBRL

Initial Input
- Number of trial: $N_{trial}$
- Rollout length: $L_{rollout}$
- Empty sample set: $(\tilde{Z}, \tilde{Y})$
- pMPC horizon: $H$
- Long-term loss: $L_d$, Immediate loss: $\ell_d$
- Awareness parameters: $\alpha_s, \alpha_t$

# Generate a scratch LGM model

# MBRL process

for $i = 1, 2, ..., N_{trial}$ do

| Step $i$ | Step $i + 1$ |
|----------|--------------|
| Read $z_i$ | Read $z_{i+1}$ |
| $u_i$ active | $u_{i+1}$ active |
| Plan $\hat{u}_i$ | Plan $\hat{u}_{i+1}$ |
| Operate $\hat{u}_i$ | Operate $\hat{u}_{i+1}$ |
| Pause $\Delta t$ | Pause $\Delta t$ |

# Ahead pMPC planning with Eq.(9)

$z_1 = \text{ObserveState}()$

$u^* = \arg\min_u L_d(z_1, H, f_d, \ell_d, \alpha_s, \alpha_t)$

for $j = 1, 2, ..., L_{rollout}$ do

$z_j = \text{ObserveState}()$

Operate($u^*(1)$)

$\hat{z}_j = [z_j, u^*]$  # Ahead pMPC planning with Eq.(9)

$\hat{z}_{j+1} = f_d(\hat{z}_j)$

$u^* = \arg\min_u L_d(\hat{z}_{j+1}, H, f_d, \ell_d, \alpha_s, \alpha_t)$

$y_j = \text{ObserveState}()$

$\hat{Z} = \{\hat{Z}, \hat{z}_j\}$, $Y = \{Y, y_j\}$  # Update LGM model

$f_d \leftarrow \text{trainLGM}(\hat{Z}, Y)$

V. Experimental Evaluation

A. Overall Definition

We conducted experiments in both simulated and real-robot arm with HEBI robotics’ hardware, Fig. 5. The simulation results demonstrate the effectiveness of contact-safe MBRL with comprehensive analysis under various awareness settings. The real-robot experiments demonstrate the real-world potential of the contact-safe MBRL with two kitchen tasks, mixing and scooping tasks.

Due to the periodic nature of robot’s rotational joint, the position and velocity are defined as: $p := [\sin \Theta^T, \cos \Theta^T]^T$, $\dot{p} := \dot{\Theta}$, where $\Theta = [\theta_1, ..., \theta_k]^T$ denotes the rotational position of a $k$-Degree-of-Freedom (DoF) robot. Also, to achieve smooth movements, joint’s velocities are used as the command-space, with a discrete acceleration $u = \Delta \dot{p} / \Delta t$ as the control-space for the pMPC. Specifically, given an optimal control signal $u_t$, the velocity commands $c_t$ are sent to the robot: $c_t = p_t + u_t$.

Both simulated and real robots tasks aims to learn the robot’s dynamics from scratch while attempting to perform the intended robotic task by tracking a provided hint, the world-space reference trajectories of performing each task. To encourage the robot to fulfill the objective, we designed the loss function to judge the robot as follow:

$$\ell_d (z) = k_s \|s^{ref} - s\|^2 + k_o \|o^{ref} - o\|^2$$  \(14\)

where $s^{ref}, o^{ref}$ are the reference position and orientation in the world-space. $k_s, k_o \in \mathbb{R}_{>0}$ are the corresponding weighting value. $s, o$ are the robot’s end-effector position and orientation in the world-space, calculated through forward kinematics with joint positions obtained from the state $z$ via Euler equation: $\Theta = \text{Re} [-i (\mu_{cos} \Theta + i \mu_{sin} \Theta)]$. Note that the design of immediate loss function $\ell_d$ does not contains any control-related term, where the uncertainty-aware control limits contribute to all behavior changes. In the followings, the weighting values are set to $k_s, k_o = 1.0$, and the tracking error is defined as $\|s^{ref} - s\|^2$.

All experiments share the following parameters: number of features $M = 65$, pMPC horizon $H = 3$, $\beta_s = 0.3$, $\gamma_t = 0.2$, $[u_{min}, u_{max}] = [-0.2, 0.2]$ rad/s.

B. Simulation Experiment

1) Simulator Setup

We setup up a particle-mixing task performed by a simple two-DoF simulated HEBI arm, with a total arm length of 0.65m and a cylinder-shaped stirring stick attached to the end-effector of the arm (Fig. 5 left). The stick is 2cm in diameter and 50cm in length, where a virtual force sensor is attached to measure the force between the arm and the stick during the mixing. A 24cm diameter round bowl is placed below the arm, filled with 900 particles in three-centimeter diameters and 0.01kg mass each. The reference trajectory follows a circulating pattern with a ten-centimeter diameter and a five-second period. The learning process contains 12 trials, 100 rollouts each with 0.1 second duration.
2) Overall Result:

This section demonstrates the tunability and effectiveness of the contact-safe MBRL in the sense of contact-safe and learning efficiency. We conducted 20 individual experiments with all combination of the following uncertainty-awareness settings: \([\alpha_s = 0, \ln \alpha_s = 6, \ln \alpha_s = 9]\) and \([\alpha_t = 0, \ln \alpha_t = 6, \ln \alpha_t = 9]\). For each combination, we calculate the followings: 1) Average tracking error at each trial over 20 experiments to demonstrate the effect onto the learning efficiency. 2) Collect the top 3%, 5%, 10% measured force on the stick and average over 20 experiments to demonstrate general effect of contact-safe MBRL in reducing the contacts’ intensiveness.

Fig. 6 demonstrates the effectiveness of the contact-safe MBRL. Being overly conservative, combinations with the setting of \(\ln \alpha_s = 9\) failed to acquire the task because the scaling factor \(K_s(\cdot)\) stays low although adequate samples. All other combinations acquired the task at a similar learning efficiency. Comparing the max measured forces, by increasing the value of \(\alpha_s\) or \(\alpha_t\), the contact-safe MBRL significantly reduced the contacts’ force. Specifically for this task, combinations of \(\{\ln \alpha_s = 6, \ln \alpha_t = 6\}\) and \(\{\ln \alpha_s = 6, \ln \alpha_t = 9\}\) acquired acquired the task at a similar learning efficiency to others while significantly reduced the contact intensiveness. This result exemplifies the tunability in uncertainty-awareness of the contact-safe MBRL.

3) Individual Result:

We selected individual results from the following settings to demonstrate the effect of the contact-safe MBRL, Fig 7, effect of translating only with settings: \(\ln \alpha_s = 6, 9\); effect of scaling only with settings: \(\ln \alpha_s = 6, 9\); effect of combination: \(\{\ln \alpha_s = 6, \ln \alpha_t = 6\}\).

In settings (1,2,3), increasing the \(\alpha_t\) effectively reduces the edge approaching velocity, Fig. 7-A. However, the significant decrease in velocity while contacting the edge reduces the effect of control limits translating (Fig. 8 upper). Consequently, without scaling, intensive contacts may occur despite its lower approaching velocity, Fig. 7-B. Next, settings (4,5) disabled the control limits translating. As shown in Fig. 6, the setting (5)’s conservative behavior failed the task acquisition. The setting (4) emphasizes the essence of the safe-characteristic (B) by its result that, without translating, the robot’s ability of reducing its momentum is limited in while uncertainty increases (Fig. 8 lower). Consequently, the setting (4) result in intensive contacts, Fig. 7-C. By enabling both scaling and translating, the setting (6) acquired the task at a much lower overall contact forces with its capability of exploring contacts safely, Fig. 7-D.

Although the setting (6) shows a relatively smaller exploration coverage, our approach does not discourage the robot from exploring new areas but encourages the robot to explore it safely. As a result, with the contact-safe MBRL, the robot can safely expand its knowledge converge until it could perform the intended robotic task.

C. Real Robot Experiment

The real-robot experiment contains two tasks: mixing task to verify the real-world performance from simulation results, and scooping task to demonstrate the potential of the proposed method at a larger scale(Fig. 5 middle, right). Both task is conducted with contact-safe control limits disabled \(\{\alpha_s = 0, \alpha_t = 0\}\), and enabled \(\{\ln \alpha_s = 5, \ln \alpha_t = 5\}\). The mixing task result is compared in following perspectives: learning efficiency, measured joint torques, environments after task acquisition. For scooping task, we only shows the environment difference after task acquired as a example of applying the contact-safe MBRL at a larger scale.

1) Mixing Task with Two-DoF Arm Configuration:

Task setup: A similar configuration in simulation is set for the real-robot mixing task with HEBI hardware. The 24cm diameter bowl measures is filled with straw cut pieces, imitating a cooking ingredient of medium viscosity. The
Fig. 7: Stick force (upper) and velocity (lower) analysis under various awareness settings through 12 trials. The figure is drawn in top view with bowl boundary (black solid line) and reference mixing trajectory (black dashed line). Color difference indicates the max value observed at each location in the bowl. (1): base-line MBRL without uncertainty-aware control limits; uncertainty-aware control limits with (2,3): scaling disabled, (4,5): translating disabled, and (6): the best combination setting.

Fig. 8: Control limits changing (upper) with and without scaling, and (lower) with and without translating.

Fig. 9: (Upper) Average tracking error and (lower) average and max(△) measured torques through trials. (Mixing task) reference trajectory follows a circulating pattern with a ten-centimeter diameter and a three-second period. The learning process contains 15 trials, 100 rollouts each with 0.1 second duration.

Results: Under intensive contacts, measured joint torques from the actuator feedbacks will also increase. Therefore, we collect joint torques during the MBRL process to compare the contacts’ intensiveness. Under the setting \{ln α_s, ln α_t = 5\}, our approach shows similar results to the simulation that contact forces are significantly reduced (Fig. 9: lower), while minimum compromises in learning efficiency (Fig. 9: upper). Also, judging the number of straws scattered in the environment during the learning process, the system with \{ln α_s, ln α_t = 5\} results in fewer straws scattered to the environment, Fig. 10, which verifies the reducing in contact intensiveness with our approach.

2) Scooping Task with Four-DoF Arm Configuration:

Task setup: The scooping task comprises a four-DoF arm with a spoon attached to the end-effector and two bowls, one is filled with straw cuts, and another is initially empty. The reference trajectory is a series of world-space spoon’s position, and orientation that scoops and transport straws. This experiment demonstrates how our approach changes the learning behavior at a larger workspace. The learning process contains 40 trials, 65 rollouts each with 0.2 second duration.

Results: Both settings achieved over six success scoops within 40 trials with a very similar pattern in reducing measured torques shown in Fig 9: lower. For the sake of spacing, here we only show the environment results. After six success scoops that transport straws to the target bowl, judging by the number of straws scattered to the environment, the contact-safe MBRL significantly reduces the intensiveness of the learning process, Fig. 10. This result verifies the behavior adjusting ability with the proposed contact-safe MBRL at a larger scale. The learning process of these two settings is shown in the attached video.
with evidence shown in the result, this work presents the contact-safe MBRL that significantly reduces contacts’ intensiveness of the learning process by utilizing the uncertainty information. Since the control limits function is independently designed apart from the learning algorithm, one can redefine such function to meet different behaviors for different purposes. Given its flexibility, and the rich information lies in uncertainties, exploring its possibilities are also one of the pivotal plans in our future road map.

VI. DISCUSSIONS AND FUTURE WORK

With evidence shown in the result, this work presents the contact-safe MBRL that significantly reduces contacts’ intensiveness of the learning process by utilizing the uncertainty information. Since the control limits function is independently designed apart from the learning algorithm, one can redefine such function to meet different behaviors for different purposes. Given its flexibility, and the rich information lies in uncertainties, exploring its possibilities are also one of the pivotal plans in our future road map.

REFERENCES

[1] M. P. Deisenroth, G. Neumann, and J. Peters, “A survey on policy search for robotics,” Foundations and Trends in Robotics, vol. 2, pp. 1–142, 2013.
[2] M. Deisenroth, C. Rasmussen, and D. Fox, “Learning to control a low-cost manipulator using data-efficient reinforcement learning,” in Robotics: Science and Systems, 2011.
[3] M. Kalakrishnan, L. Righetti, P. Pastor, and S. Schaal, “Learning force control policies for compliant manipulation,” in 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 4639–4644, 2011.
[4] P. Pastor, M. Kalakrishnan, S. Chitta, E. Theodorou, and S. Schaal, “Skill learning and task outcome prediction for manipulation,” in 2011 IEEE International Conference on Robotics and Automation, pp. 3828–3834, 2011.
[5] I. Misiołek, Y. Karasjannidis, J. A. Stork, and D. Kräger, “Data-driven model predictive control for the contact-rich task of food cutting,” in 19th IEEE-RAS International Conference on Humanoid Robots, pp. 24–250, 2019.
[6] P. F. S. A. Khader, H. Yin and D. Kräger, “Probabilistic model learning and long-term prediction for contact-rich manipulation tasks,” CoRR, vol. abs/1909.04915, 2019.
[7] D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. M. Scokaert, “Constrained model predictive control: Stability and optimality,” Automatica, vol. 36, no. 6, pp. 789–814, 2000.
[8] G. Cao, E. M. Lai, and F. Alani, “Gaussian process model predictive control of an unmanned quadrotor,” Journal of Intelligent and Robotic Systems, vol. 88, no. 1, pp. 147–162, 2017.
[9] R. N. Carleton, “Into the unknown: A review and synthesis of contemporary models involving uncertainty,” Journal of Anxiety Disorders, vol. 39, pp. 30 – 43, 2016.
[10] R. N. Carleton, “Fear of the unknown: One fear to rule them all?,” Journal of Anxiety Disorders, vol. 41, pp. 5 – 21, 2016.
[11] S. Calinon, I. Sardellitti, and D. G. Caldwell, “Learning-based control strategy for safe human-robot interaction exploiting task and robot redundancies,” in 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 249–254, 2010.
[12] M. P. Deisenroth and C. E. Rasmussen, “Pilco: A model-based and data-efficient approach to policy search,” in Proceedings of the 28th International Conference on International Conference on Machine Learning, pp. 465–472, 2011.
[13] H. Schäfler and U. Ledewitz, Geometric Optimal Control: Theory, Methods and Examples, vol. 53, 2012.
[14] S. Kamthe and M. P. Deisenroth, “Data-efficient reinforcement learning with probabilistic model predictive control,” in International Conference on Artificial Intelligence and Statistics, vol. 84, pp. 1701–1710, 2018.
[15] Y. Cui, Y. Cui, and T. Matsubara, “Sample-and-computation-efficient probabilistic model predictive control with random features,” in 2020 IEEE International Conference on Robotics and Automation, pp. 307–313, 2020.
[16] L. S. Pontryagin, E. F. Mishchenko, V. G. Boltyan-ski, and R. V. Gamkrelidze, The Mathematical Theory of Optimal Processes. Wiley, 1962.
[17] K. Lynch and F. Park, Modern Robotics: Mechanics, Planning, and Control, Cambridge University Press, 2017.
[18] G. Schoettler, A. Nair, J. Luo, S. Bahl, J. A. Ojea, E. Solowjow, and S. Levine, “Deep reinforcement learning for industrial insertion tasks with visual inputs and natural rewards,” CoRR, vol. abs/1906.05841, 2019.
[19] L. Johannsmeyer, M. Gerchow, and S. Haddadin, “A framework for robot manipulation: Skill formalism, meta learning and adaptive control,” in 2019 IEEE International Conference on Robotics and Automation, pp. 5844–5850, 2019.
[20] C. Beltran, D. Petit, I. G. Ramirez-Alpizar, T. Nishi, S. Kikuchi, T. Matsubara, and K. Harada, “Learning force control for contact-rich manipulation tasks with rigid position-controlled robots,” IEEE Robotics and Automation Letters, vol. 5, pp. 5709 – 5716, 2020.
[21] F. Wimshofer, P. Schmitt, G. von Wichert, and W. Burgard, “Controlling contact-rich manipulation under partial observability,” in Robotics: Science and Systems, 2020.
[22] S. Levine, N. Wagener, and P. Abbeel, “Learning contact-rich manipulation skills with guided policy search,” in IEEE International Conference on Robotics and Automation, pp. 156–163, 2015.
[23] K. Deb, Multi-Objective Optimization Using Evolutionary Algorithms. John Wiley & Sons, Inc., 2001.
[24] T. Miyamoto, H. Sasaki, and T. Matsubara, “Exploiting visual-outer shape for tactile-inner shape estimation of objects covered with soft materials,” IEEE Robotics and Automation Letters, vol. 5, no. 4, pp. 6278–6285, 2020.
[25] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, “Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor,” in Proceedings of the 35th International Conference on Machine Learning (J. G. Dy and A. Krause, eds.), pp. 1856–1865, 2018.
[26] Y. Cui, S. Osaki, and T. Matsubara, “Reinforcement learning ship autopilot: Sample-efficient and model predictive control-based approach,” in 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2868–2875, 2019.
[27] M. A. Lee, C. Florensa, J. Tremblay, N. Ratliff, A. Garg, F. Ramos, and D. Fox, “Guided uncertainty-aware policy optimization: Combining learning and model-based strategies for sample-efficient policy learning,” in 2020 IEEE International Conference on Robotics and Automation, pp. 7505–7512, 2020.
[28] C. K. I. Williams, “Prediction with gaussian processes: From linear regression to linear prediction and beyond,” in Learning in Graphical Models, pp. 599–621, Springer Netherlands, 1998.
[29] Q. Le, T. Sariös, and A. Smola, “Fastfood: Approximating kernel expansions in loglinear time,” in Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28, pp. 244–252, 2013.
[30] W. Rudin, Fourier Analysis on Groups. Wiley Classics Library, reprint edition, 1994.
[31] C. Rasmussen and C. Williams, Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press, 2006.
[32] C. M. Bishop, Pattern Recognition and Machine Learning. Springer, 2006.
[33] Y. Cui, S. Osaki, and T. Matsubara, “Autonomous boat driving system using sample-efficient model predictive control-based reinforcement learning approach,” Journal of Field Robotics, 2020.