Context-Aware Safe Reinforcement Learning
for Non-Stationary Environments

Baiming Chen, Zuxin Liu, Jiacheng Zhu, Mengdi Xu, Wenhao Ding, Ding Zhao

Abstract—**Safety is a critical concern when deploying reinforcement learning agents for realistic tasks. Recently, safe reinforcement learning algorithms have been developed to optimize the agent's performance while avoiding violations of safety constraints. However, few studies have addressed the non-stationary disturbances in the environments, which may cause catastrophic outcomes. In this paper, we propose the context-aware safe reinforcement learning (CASRL) method, a meta-learning framework to realize safe adaptation in non-stationary environments. We use a probabilistic latent variable model to achieve fast inference of the posterior environment transition distribution given the context data. Safety constraints are then evaluated with uncertainty-aware trajectory sampling. The high cost of safety violations leads to the rareness of unsafe records in the dataset. We address this issue by enabling prioritized sampling during model training and formulating prior safety constraints with domain knowledge during constrained planning. The algorithm is evaluated in realistic safety-critical environments with non-stationary disturbances. Results show that the proposed algorithm significantly outperforms existing baselines in terms of safety and robustness.**

I. INTRODUCTION

Reinforcement learning (RL) is a promising way to solve sequential decision-making tasks. For example, RL has shown superhuman performance in competitive games like Go [1] and Starcraft [2]. RL has also been used for the control of complex robotic systems [3], [4] such as legged robots [5]. However, most well-known RL algorithms [6], [7], [8] do not consider safety constraints during exploration. Moreover, they are usually not adaptive to non-stationary disturbances, which are common in many realistic safety-critical applications [9]. These two weaknesses of current RL algorithms need to be addressed before their deployment in safety-critical environments.

Several recent studies have been proposed to address the lack of **safety** [10], [11], [12], [13] and the lack of **adaptability** [14], [15], [16] issues of RL algorithms, respectively. However, the above two issues are entangled in realistic applications, because the environment disturbances may change the system dynamics and affect the region of safety. In other words, disturbances may cause unexpected safety violations if not properly handled. A typical example is shown in Fig. 1, where a healthcare robot is trying to deliver the medicine

*(Corresponding author: Ding Zhao)

Baiming Chen is with Tsinghua University, Beijing, China. This work was done during his visit in Carnegie Mellon University (e-mail: cbm17@mails.tsinghua.edu.cn)

Zuxin Liu, Jiacheng Zhu, Mengdi Xu, Wenhao Ding and Ding Zhao are with the Department of Mechanical Engineering, Carnegie Mellon University, USA (e-mail: {zuxinl, jzhu4, mengdixu, wenhaod, dingzhao}@andrew.cmu.edu)
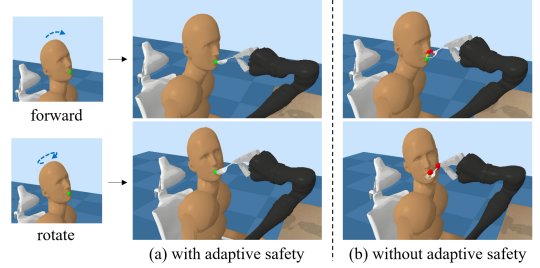
Fig. 1: Healthcare environment with and without adaptive safety. Red dots indicate direct contacts between the robot and the patient which should be avoided.

(or food) to the patient while avoiding any direct contact. The disturbance in this environment mainly comes from the patient's movements. To safely finish the delivery, the robot must be able to quickly identify the patient's moving preference and adaptively generate safe control decisions. To the best of our knowledge, there hasn't been a general framework or a complete algorithm to fully address this entangled problem.

In this paper, we propose the context-aware safe reinforcement learning (CASRL) framework to realize safe adaptation in non-stationary environments and resolve the above entangled problem. Our major contribution is threefold:

1) **Fast adaptation**. We study this problem under the model-based RL framework for sample efficiency. Unlike previous models that predict the next state only based on the current state and action, we use a context-aware latent variable model to infer the disturbance of the non-stationary environment based on the historical transition data, allowing task-agnostic adaptation.

2) **Risk-averse control**. We achieve risk-averse decision making with constrained model predictive control. Constraints are used for guarantees of safety in uncertain environments. To improve exploration safety in the early stage of training, we incorporate domain knowledge to make conservative decisions with prior models. We also enable prioritized sampling of rare unsafe data during the model training to alleviate the data imbalance problem in safety-critical environments. Combined with a context-aware probabilistic model, this control regime can realize safe adaptation in non-stationary environments and resolve the aforementioned entangled problem.

3) **Extensive evaluation**. We conduct experiments in a toy example and a realistic high-dimensional environ-

ment with non-stationary disturbances. Results show that the proposed method can (i) realize fast adaptation for safe control in unseen environments, (ii) scale to high-dimensional tasks, and (iii) outperform existing approaches in terms of safety and robustness.

II. RELATED WORK

Safe reinforcement learning has attracted long-term interest in the RL community [17]. The Constrained Markov Decision Processes (CMDPs) [18] is often used to model the safe RL problem, where the agent aims to maximize its cumulative reward while satisfying certain safety constraints. Several approaches, such as the Lagrangian method [19] and constrained policy optimization [10], [20], have been proposed to solve CMDPs. Gaussian Processes (GPs) have also been used to approximate the dynamics of the environment for safe exploration [21], [22]. Particularly, Wachi and Sui [23] discussed the situation where the safety boundary is unknown. However, most existing safe RL methods assume a consistent environment and cannot deal with time-varying disturbances. In contrast, our method aims to realize safe control in non-stationary environments, which is more realistic for safety-critical applications.

Robust adversarial learning addresses the environment disturbance problem by formulating a two-player zero-sum game between the agent and the disturbance [24], [25], [26]. However, the robust policies trained in this way may overfit to the worst-case scenario, so the performance is not guaranteed in other cases [27].

Meta-learning for RL has recently been developed to realize adaptive control in non-stationary environments [28], [14], [29], [15], [16], [30]. Since unsafe data are particularly rare in safety-critical environments, we focus on model-based methods for sample efficiency [8]. Sæmundsson et al. [29] proposed to use Gaussian Processes to represent dynamics models, which may suffer from poor scalability as the dimension and the amount of data increases. Nagabandi et al. [15] integrated model-agnostic meta-learning (MAML) [14] with model-based RL. The dynamics model is represented by a neural network that uses a meta-learned initialization and is quickly updated with the latest data batch. However, the uncertainty is not estimated by the model, and we show that this may degrade the performance. Later studies from Xu et al. [16] and Nagabandi et al. [3] achieved online continual learning with streaming data by maintaining a mixture of meta-trained dynamics models. These approaches may suffer from the model explosion in complex environments where the potential number of dynamics type is large. We overcome this issue by constructing a probabilistic latent variable model that learns a continuous mapping from the disturbance space to the latent space.

Neural Processes (NPs) [31] have been proposed for few-shot regression by learning to map a context set of input-output observations to a distribution of regression functions. Comparing to the Gaussian processes, NPs have the advantage of efficient data-fitting with linear complexity in the size of context pairs and can learn conditional distributions

with a latent space. A later study [32] proposed Attentive Neural Processes (ANPs) by incorporating attention into NPs to alleviate the underfitting problem and improve the regression performance. NP-based models have shown great performance in function regression [33], image reconstruction [32], and point-cloud modeling [34]. As probabilistic latent variable models, ANPs naturally enable continual online learning in continuously parameterized environments. In this paper, we will show how to incorporate ANPs for dynamics prediction and safety constraint estimation.

The rest of the paper is organized as follows. In Sec. III, we formulate the safety-critical problem that we aim to solve in this paper. In Sec. IV, we show the inference process of unknown environment disturbances with a latent variable model. In Sec. V, we show how to perform safe adaptation with a sampling-based model-predictive controller. The experiment results and discussions are presented in Sec. VI.

III. PROBLEM STATEMENT

We consider non-stationary Markov Decision Processes (MDPs) with safe constraints. An MDP is defined as a tuple $(\mathcal{S}, \mathcal{A}, f, r, \gamma, \rho_0)$ where $\mathcal{S}$ denotes the state space, $\mathcal{A}$ denotes the action space, $f(s'|s, a)$ is the transition distribution of the environment dynamics that takes into the current state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$, and outputs the distribution of the next state $s' \in \mathcal{S}$. $r(s, a)$ is the reward function, γ is the reward discount factor, and ρ_0 is the distribution of the initial state. To simulate the disturbances in real-world environments, we consider non-stationary MDPs where the transition dynamics $f(s'|s, a, \theta)$ depends on certain hidden parameters $\theta \sim \mathcal{T}$, where $\mathcal{T}$ denotes the distributions of environments parameters. For simplicity, we assume that the environment is episodically consistent - the change of f only happens at the beginning of each episode. This setting is commonly used in related papers and can be easily generalized to other consistent time-horizons.

Denote a safe state set by $\mathcal{S}_{safe}$ and a safe action set by $\mathcal{A}_{safe}$. The goal of safe RL is to find the optimal action sequence $a_{0:T}$ to maximize the discounted accumulated reward $\sum_{t=0}^{\tau} \gamma^t r(s_t, a_t)$, without violating the safety constraints (i.e., keeping $s_t \in \mathcal{S}_{safe}$ and $a_t \in \mathcal{A}_{safe}$ for every time step t). γ is a discount factor and τ is the task horizon. Throughout this paper, we assume $\mathcal{S}_{safe}$ and $\mathcal{A}_{safe}$ are known a *priori*.

IV. CONTEXT-AWARE MODEL INFERENCE

We address the proposed problem under the model-based RL framework, where the tasks are solved by learning a dynamics model $\tilde{f}(s'|s, a)$ to approximate the ground-truth environment dynamics $f(s'|s, a)$. However, when the environment dynamics f is non-stationary, $\tilde{f}(s'|s, a)$ may fail to make accurate predictions since some hidden features of the environment are not identified. To handle this problem, we propose to learn a context-aware model $\tilde{f}(s'|s, a, C)$ that performs state predictions based not only on the current state s and action a but also on the *contexts* C - the historical data collected in the current episode. In this way, the hidden
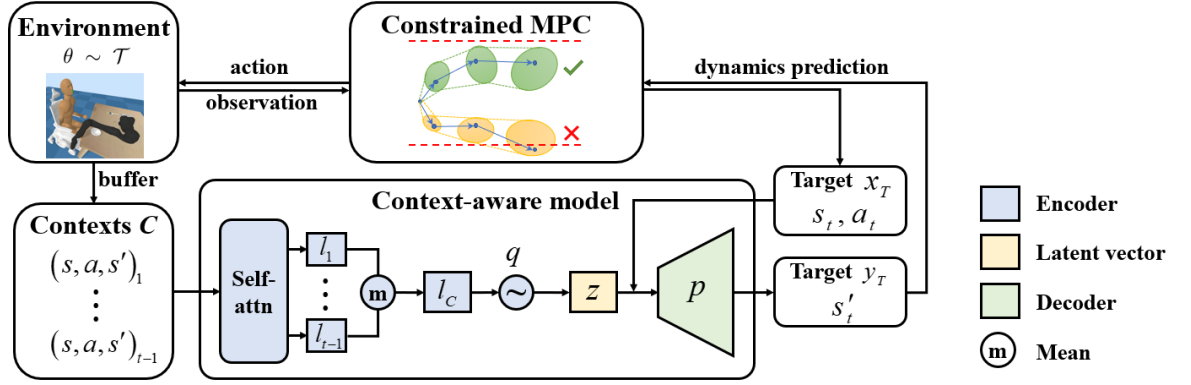
Fig. 2: The flow of the proposed context-aware safe reinforcement learning (CASRL) framework. A context-aware model is used to perform conditional dynamics predictions based on the context data.

information of the environment is first inferred from C, and then the posterior distribution of the next state s' is calculated.

To incorporate domain knowledge for adaptive learning, we divide the dynamics model $\tilde{f}(s'|s, a, C)$ into two parts:

$$s' := s'_h + s'_g, \tag{1a}$$

$$\text{with} \quad s'_h \sim h(\cdot|s, a), \tag{1b}$$

$$s'_g \sim g(\cdot|s, a, C). \tag{1c}$$

The model h in Eq. (1b) is referred to as the *prior model*. Such model can be obtained by leveraging domain knowledge without necessarily interacting with the environment, e.g., training the dynamics model in a simulator [8] or using first principles modeling [35]. However, the drawback is that they are usually context-unaware.

The model g in Eq. (1c) is called the *disturbance model* (or the error model). It represents the error between the prior model h and the overall dynamics model $\tilde{f}$. It is the model we aim to learn by interacting with the target non-stationary environment. The disturbance model is context-aware and should be able to capture the hidden information of the environment based on the contexts C. To achieve that, the disturbance model g should have the following properties:

- Flexibility: g should be able to condition on arbitrary number of contexts to make predictions.
- Uncertainty awareness: g should estimate the uncertainty in its predictions to balance exploration and exploitation.
- Scalability: g should be able to scale to high-dimensional environments.

In this paper, we use an Attentive Neural Process (ANP) [32] to represent the disturbance dynamics model g for its desirable properties and implementation simplicity. The ANP model is defined as a (infinite) family of conditional distributions, in which an arbitrary number of observed input-output *contexts* $(x_C, y_C) := (x_i, y_i)_{i \in C}$ is used to model an arbitrary number of input-output *targets* $(x_T, y_T) := (x_i, y_i)_{i \in T}$, where C denotes a set of observed points and T denotes a set of unobserved points (the output

y_T is unknown). The ANP transforms the original conditional likelihood to a hierarchical inference structure:

$$g(y_T|x_T, x_C, y_C) = \int p(y_T|x_T, z) \, q(z|l_C) \, dz \tag{2}$$

where z is a global latent vector describing uncertainty in the predictions of y_T for given observations (x_C, y_C), and is modeled by a factorized Gaussian parameterized by $l_C := l(x_C, y_C)$, with l being a deterministic function that aggregates (x_C, y_C) into a fixed dimensional representation. In ANP, l consists of a multilayer perceptron (MLP), self-attentions, and a mean aggregation layer to produce permutation-invariant representations.

For dynamics prediction, the input x is the state-action pair (s, a), and the output y is the state at the next time step s'. At time t, the contexts $(x_C, y_C) = (s_i, a_i, s'_i)_{i \in [1:t-1]}$ contain the state-action information of the previous time steps, the target input $x_T = (s_t, a_t)$ is the current state-action pair, and we aim to predict the target output $y_T = s'_t$ that represents the next state. The flow of using context-aware model for model-based RL is shown in Fig. 2. A constrained MPC controller is used for safe planning and will be introduced in the next section.

The training of ANP is based on the amortized variational inference. The parameters of the encoders and the decoder are updated by maximizing the following evidence lower bound (ELBO) with the reparametrization trick [36]:

$$\log g(y_T|x_T, x_C, y_C) \geq$$
$$\mathbb{E}_{q(z|l_T)} \left[\log g(y_T|x_T, z) \right] - D_{\text{KL}} \left(q(z|l_T) \parallel q(z|l_C) \right). \tag{3}$$

where $l_T := l(x_T, y_T)$, with l being a deterministic function introduced before. The training objective of ANP can be interpreted as improving the prediction accuracy on the targets while regularizing the Kullback–Leibler divergence between the latent encoding of the contexts and the targets.

The contexts and the targets are randomly sampled from a replay buffer that stores transition data from the same disturbance dynamics. However, the rareness of unsafe data may lead to low prediction accuracy in the unsafe state region. To alleviate this issue, inspired by [37], we enable prioritized experience sampling during model training - to

train the context-aware model with a certain data batch, the unsafe data in this data batch are first added into the target set T, and then other safe data are uniformly sampled and appended to C and T. We found that this trick can effectively increase the prediction accuracy in the unsafe region, which is discussed in Sec. VI-C.4.

V. SAFE ADAPTATION WITH MPC

We formulate the safe adaptation as a constrained nonlinear optimization problem:

$$\max_{a_{0:\tau}} \quad \sum_{t=0}^{\tau} r(s_t, a_t) \tag{4a}$$

$$\text{s.t.} \quad a_t \in \mathcal{A}_{safe} \tag{4b}$$

$$s_{t+1} \sim \tilde{f}(\cdot|s_t, a_t, C) \tag{4c}$$

$$\Pr(s_t \notin \mathcal{S}_{safe}) \leq \delta \tag{4d}$$

$$\hat{s}_{t+1} \sim h(\cdot|\hat{s}_t, a_t) \tag{4e}$$

$$\Pr(\hat{s}_t \notin \mathcal{S}_{safe}) \leq \delta \tag{4f}$$

$$\text{for} \quad t = 0, \ldots, \tau$$

Eq. (4a) shows that the objective is to maximize the cumulative reward, Eq. (4b) represents the safety constraint on actions, and Eq. (4c, 4d) define the safety constraint on the states s_t that predicted by the learned model $\tilde{f}$. Eq. (4a)-(4d) form the general problem of safe RL in most previous literature [21]. However, with the non-stationary environment disturbances, the learning process of the prediction model $\tilde{f}$ may be unstable, and it is difficult for the agent to keep safe when $\tilde{f}$ is not accurate. To alleviate this problem, we formulate the prior safety constraint shown in Eq. (4e, 4f), where a sequence of auxiliary states $\hat{s}_t$ is predicted only with the prior model h, and the high-probability safety constraint is applied to it ($\hat{s}_0 = s_0$). Though not accurate, the prior safety constraint provides extra protection for the agent based on the static prior model h. Applying the prior safety constraint is an effective way to incorporate domain knowledge to improve safe learning, especially when the unsafe data are expensive to obtain. Experiment results show that it can effectively reduce the safety violation rate especially in the early stage of training (Sec. VI-C.1).

Direct solving the optimization problem Eq. (4) is intractable since $\tilde{f}$ is a high-dimensional nonlinear stochastic function. Previous work has used approximated uncertainty propagation techniques like sigma-point transform [38] and Taylor expansion [21] to model the state distribution as a single Gaussian distribution, and then solve Eq. (4) with nonlinear solvers such as the IPOPT [39]. However, Deisenroth et al. [40] showed that the Gaussian moment matching could corrupt after long-term propagation due to the multi-modal distribution of states, inducing huge prediction errors. Also, IPOPT cannot provide an alternative plan if no solution for Eq. (4) is found in limited time.

In this paper, we propose to solve Eq. (4) with a sampling-based model-predictive control (MPC) approach. We use MPC for its implementation simplicity, time flexibility, and risk aversion. Also, this sampling-based method makes no

---

Algorithm 1 Trajectory sampling

procedure TRAJSAMPLING(A, h, g, C, t_0)
    for SamplingTime $= 1, N$ **do**
        for $t = t_0, t_0 + \tau_p$ **do**
            $s_{th} \sim h(\cdot|s_{t-1}, a_{t-1})$
            $s_{tg} \sim g(\cdot|s_{t-1}, a_{t-1}, C)$
            $s_t = s_{th} + s_{tg}$
    return $\{s_{t_0:t_0+\tau}\}_{1:N}$

---

assumptions on the pattern of state distributions. Denoting the planning horizon with τ_p, we first define the augmented objective function for an action sequence $A = a_{t_0:t_0+\tau_p}$ as:

$$\bar{R}(A) := \sum_{t=t_0}^{t_0+\tau_p} [r(s_t, a_t) - \lambda[(\mathbb{1}(\Pr(s_t \notin \mathcal{S}_{safe}) > \delta) + \mathbb{1}(\Pr(\hat{s}_t \notin \mathcal{S}_{safe}) > \delta) + \mathbb{1}(a_t \notin \mathcal{A}_{safe}))]] \tag{5}$$

where $\mathbb{1}(Z)$ is the indicator function that returns 1 if Z is true, otherwise 0. s_t and $\hat{s}_t$ are the state particles defined in Eq. (4) and are produced by the trajectory sampling procedure where the uncertainties are propagated (Alg 1). λ serves as the Lagrangian multiplier of the dual problem of Eq. (4). In this paper, we regard λ as a fixed hyperparameter and make it sufficiently large

$$\lambda \geq \max(|r|) * \tau \tag{6}$$

so that the augmented performance is monotonically decreasing w.r.t. the safety violation number. Considering the uncertainty in the probabilistic model, we evaluate A with the *Conditional Value at Risk (CVaR)* [41] of $\bar{R}(A)$ to make the solutions risk-averse:

$$\text{CVaR}_\alpha(\bar{R}(A)) = \mathbb{E}\left[\bar{R}(A)|\bar{R}(A) \leq \nu_\alpha(\bar{R}(A))\right] \tag{7}$$

where $\alpha \in (0, 1)$ and ν_α is the α-quantile of the distribution of $\bar{R}(A)$. In other words, we prefer action sequences with higher CVaR. We then take the first action in the most preferred action sequence and execute it. Instead of uniformly sampling A every time, we utilize the *Cross-Entropy Method (CEM)* as suggested in [8] to keep the historical information. The complete algorithm along with the model-learning part is shown in Alg. 2.

VI. EXPERIMENT

For the evaluation of the proposed algorithm, we aim to answer the following questions through empirical experiments: can CASRL 1) adapt faster to unseen environments with a stream of non-stationary data than existing approaches? 2) reduce the safety violation rate with prior safety constraints? 3) scale to high-dimensional tasks?

A. Environments

To answer the above questions, we test CASRL in two continuously-parameterized non-stationary environments with safety constraints. The setup of the environments (Fig. 3) is introduced below.

Algorithm 2 Context-Aware Safe Reinforcement Learning (CASRL)

Input: prior model h, state safe set $\mathcal{X}_{safe}$, action safe set $\mathcal{A}_{safe}$, task distribution $\mathcal{T}$
Output: disturbance model g, episodic replay buffer R
$\tilde{g} \leftarrow g_0$, $R \leftarrow \{\}$ $\triangleright$ Initialize the disturbance model and the replay buffer
for Episode = 1, M **do**
    $p \sim \mathcal{T}$, $C \leftarrow \{\}$, reset CEM($\cdot$), get s_0 $\triangleright$ Environment sampling and episode initialization
    for $t = 1, \tau$ **do**
        for $A \sim$ CEM($\cdot$) **do** $\triangleright$ Sampling action sequences
            $s_{t:t+\tau_p} = \text{TRAJSAMPLING}(A, h, g, C, s_{t-1}, t)$ $\triangleright$ State propagation in the learned model
            $\hat{s}_{t:t+\tau_p} = \text{TRAJSAMPLING}(A, h, 0, C, s_{t-1}, t)$ $\triangleright$ State propagation in the prior model
            $A^* = \arg\max_A \text{CVaR}_\alpha \left(\bar{R}(A) \right)$ $\triangleright$ The optimal action sequence is selected based on the CVaR
            Update CEM($\cdot$)
        Execute a_t^*, get s_{t+1} $\triangleright$ a_t^* is the first element of A^*
        $C \leftarrow C \cup (s_t, a_t^*, s_{t+1})$ $\triangleright$ Record context
    $R \leftarrow R \cup C$ $\triangleright$ Update the episodic replay buffer
    Update g by maximizing the ELBO in Eq. (3) with R $\triangleright$ Model learning

---

(a) cart-pole swingup   (b) healthcare feeding robot

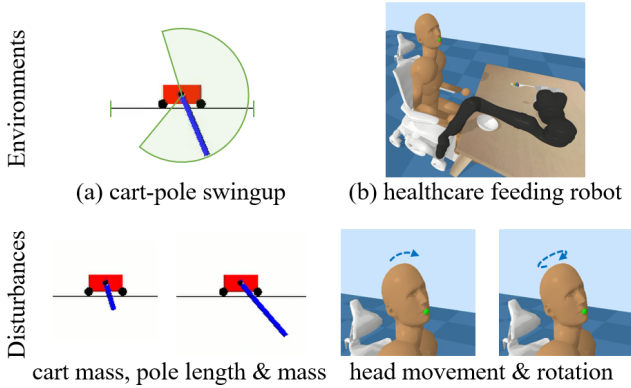cart mass, pole length & mass   head movement & rotation

Fig. 3: Tasks with non-stationary disturbances and safety constraints.

- **cart-pole.** ($\mathcal{S} \subseteq \mathbb{R}^4$, $\mathcal{A} \subseteq \mathbb{R}^1$) This is the cart-pole swingup experiment proposed in [29]. The goal is to swing the pole upright by applying force on the cart while keeping the cart close to the center of the rail. We add constraints on the pole angle $\theta \in [-10°, 225°]$ so that the pole should be swung up from the right side without too much overshoot. We make the task non-stationary by changing the pole length l, the pole mass p_m, and the cart mass c_m at the beginning of each episode. The observation includes the position x and velocity $\dot{x}$ of the cart, as well as the angle θ and angular velocity $\dot{\theta}$ of the pole. The reward function is $r = \exp\left(-\frac{(x - l\sin\theta)^2 + (l - l\cos\theta)^2}{l^2}\right)$ and the highest reward $r = 1$ is acquired when the cart is at the center of the rail ($x = 0$) and the pole is upright ($\theta = 0$). The simulation frequency is 20 Hz.
- **healthcare feeding robot.** ($\mathcal{S} \subseteq \mathbb{R}^{23}$, $\mathcal{A} \subseteq \mathbb{R}^7$) The environment is provided by [42]. The goal is to deliver the medicines to the patient's mouth with a control arm. To keep safe, there should be no direct contact between the patient and the robot. In each episode, the patient

moves forward and rotates his head in 4 degree-of-freedom with randomly sampled speeds ($a_f, a_\theta, a_\phi, a_\psi$), which is the disturbance we designed to simulate different preferences. This is a relatively high-dimensional environment and is used to test the scalability of the algorithms. The observation includes the position of the robot joints and the spoon, as well as the position and orientation of the human head. The reward function has three parts: $r = r_{dis} + r_{med} + r_{act}$, where r_{dis} penalizes the distance between the spoon and the target position, r_{med} is a large positive value if medicine particles are successfully delivered or a large negative value if they are spilled, and r_{act} penalizes the magnitude of the control input. The simulation frequency is 10 Hz.

B. Baselines

We compare our method with the following baselines:
- **Projection-Based Constrained Policy Optimization (PCPO):** A projection-based safe RL algorithm [43]. The learned policy is projected to the safe region during training.
- **Probabilistic Ensemble and Trajectory Sampling (PETS):** To evaluate the importance of context-aware adaptation, we compare to PETS [8], a state-of-the-art model-based RL approach.
- **Model-Agnostic Meta-Learning (MAML):** We use the gradient-based MAML [14], [15] to learn the dynamics of the non-stationary environments. The dynamics model is represented by a neural network which is initialized from a pre-trained meta-model and updated online with the nearest context data. [1]
- **CASRL without prior safety constraint:** To show whether the prior safety constraint can effectively reduce the safety violation rate, we add another baseline that follows the same structure of CASRL but does not apply the prior safety constraint.

---

[1] We used a publicly available implementation at https://github.com/iclavera/learning_to_adapt.

Each algorithm (including the proposed method) is first pre-trained in non-safety-critical simulators without any disturbances ($\mathcal{T}_{pre}$) to learn the prior model h, where the safety constraints are not applied so that we have enough data from both safe and unsafe regions. We then use these initialized models to safely adapt in disturbance spaces $\mathcal{T}_{adapt}$ to learn the disturbance model g, with constraints applied. As introduced in Sec. III, we re-sample the parameters of the environments from $\mathcal{T}_{adapt}$ at the beginning of each episode. The results will reflect the adaptability of the tested algorithms. $\mathcal{T}_{pre}$ and $\mathcal{T}_{adapt}$ used in the experiments are shown in Table I.

TABLE I: Disturbance Space. $\mathcal{U}(\cdot)$ denotes uniform distribution.

Environment	$\mathcal{T}_{pre}$	$\mathcal{T}_{adapt}$	Unit
cart-pole	$l = 0.6$	$l \sim \mathcal{U}[0.2, 1.0]$	m
	$p_m = 0.6$	$p_m \sim \mathcal{U}[0.2, 1.0]$	kg
	$c_m = 0.6$	$c_m \sim \mathcal{U}[0.2, 1.0]$	kg
healthcare	$a_f = 0$	$a_f \sim \mathcal{U}[-1.0, 1.0]$	°/s
	$a_\theta = 0$	$a_\theta \sim \mathcal{U}[-2.0, 2.0]$	°/s
	$a_\phi = 0$	$a_\phi \sim \mathcal{U}[-2.0, 2.0]$	°/s
	$a_\psi = 0$	$a_\psi \sim \mathcal{U}[-2.0, 2.0]$	°/s

In the implementation, we use a hidden size of $[128, 128]$ for all MLP networks. The latent dimension is 8 for the deterministic encoder and latent encoder in the ANP model for both experiments. The planning horizon τ is set to be 20. Each experiment was run with 10 random seeds. We make the controller risk-averse by setting $\delta = 0$ in Equ. 5. All hyperparameters are fine-tuned manually and are provided in our submitted code base.

C. Result Analysis

1) During Adaptive Training: The average returns and safety violation rates during adaptive training are shown in Fig. 4. The violation rate represents the proportion of safety violation time steps in the whole episode. For PCPO, we only plot the highest average performance after its convergence since it requires a lot more samples to train than other model-based methods. It is shown that the performance of PCPO is limited since it cannot deal with non-stationary environment disturbances. Though PETS outperforms other methods in most environments during the early stage of training, it fails to continue improving due to the lack of adaptability in non-stationary environments. The proposed approach, CASRL, outperforms MAML in both average returns and safety violation rates, especially in the healthcare environment. There are two possible reasons. One is that the adaptation of MAML relies on online training of a high-dimensional neural-network model in each step, which is very sensitive to the learning rate and could be unstable in high-dimensional spaces. On the other hand, CASRL only performs online inference. The other possible reason is that MAML cannot model the uncertainties in the environment, which is accomplished by CASRL with a probabilistic latent variable model.
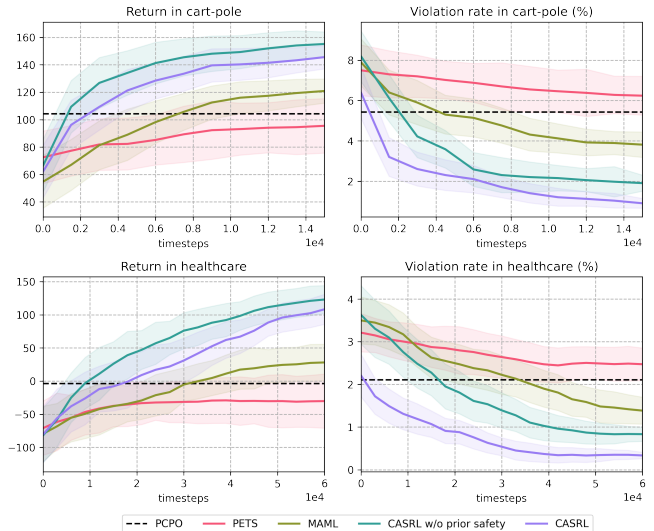

Fig. 4: Return and violation rate during adaptive training. The proposed method CASRL greatly reduces safety violation rate while outperforming MAML in average return.

We can also observe that the prior safety constraint can significantly reduce the violation rate with minimal performance degradation.

2) After Adaptive Training: We evaluate the performance of models after adaptive training by experiment in the whole disturbance space $\mathcal{T}_{adapt}$ (Tab. I). The results of average returns and safety violation rates in cartpole-swingup and healthcare are shown as heatmaps in Fig. 5. It is interesting to observe that different constraint functions can lead to different patterns of heatmaps. In the cartpole-swingup environment, most constraint-violation cases concentrate at the corners of the disturbance space (Fig. 5b) because the dynamics models in the corners are the most different from the center. In the healthcare environment, however, most constraint-violation cases take place when the human head has a high velocity of forward movement (Fig. 5d), which is reasonable since forward movement decreases the distance between the human head and the robot, increasing the risk of direct contact. Among the methods tested, CASRL shows great robustness and adaptability to disturbances compared to other baselines.

3) Effect of pre-training: The pre-training phase is essential for CASRL. The pre-trained prior model h not only provides a start point for adaptive learning but also forms the prior safety constraint that improves the safety of the learning process. To show this, we compare the performance of CASRL with and without pre-training in Fig. 6. MAML provides a baseline. It is clearly shown that the pre-training phase significantly benefits the learning process, especially for CASRL.

For the healthcare experiment, the violation rate experienced a big jump in the early stage of training for both methods. The reason is that the robot needs to learn to control its arm before it can approach the patient and possibly violate the safety constraint.

(a) Return in cart-pole

(b) Violation rate (%) in cart-pole

(c) Return in healthcare

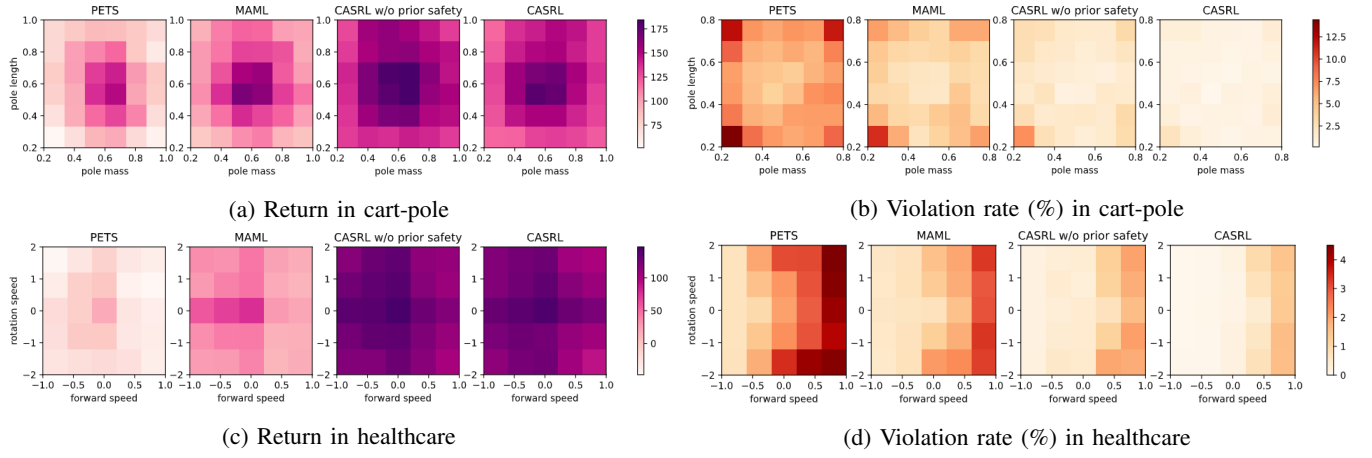(d) Violation rate (%) in healthcare

Fig. 5: Return and violation rate after adaptive training in cart-pole and healthcare environments.

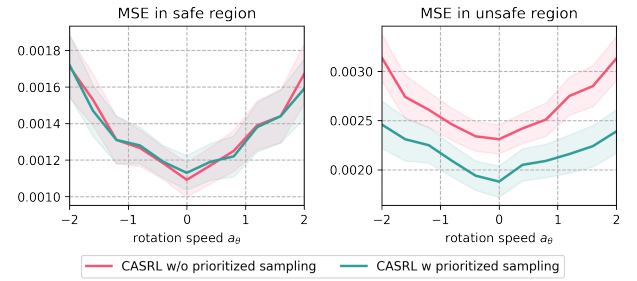Fig. 6: Comparison of CASRL and MAML with and without the pre-training phase.

Fig. 7: The MSE of single-step dynamics predictions by CASRL in healthcare environment. The prediction accuracy in the unsafe region is improved by prioritized sampling.

4) Effect of prioritized sampling: We evaluate the effectiveness of prioritized sampling by comparing the mean square error (MSE) of dynamics predictions in safe and unsafe regions. The results are shown in Fig. 7. The prediction accuracy in the unsafe state region is improved by prioritized sampling, while the performance in the safe state region is not influenced. The reason could be that without prioritized sampling, the model is biased towards the safe data due to the rareness of the unsafe samples.

VII. CONCLUSION

In this paper, we propose the context-aware safe reinforcement learning (CASRL) method as a meta-learning framework to realize safe adaptation in non-stationary environments. The non-stationary disturbances are identified with a probabilistic latent variable model by online Bayesian inference. A risk-averse model-predictive controller is used for safe planning with uncertainties, where we incorporate prior safety constraints to enable fast adaptation with prior knowledge. We also utilize prioritized sampling of unsafe data to alleviate the data imbalance in safety-critical environments. The algorithm is evaluated in both toy and realistic high-dimensional environments. Results show that CASRL significantly outperforms existing baselines in terms of safety and robustness.

Although CASRL is potentially beneficial for RL applications in safety-critical tasks, it may have its limitations. For example, the disturbance space could be much larger if we use image inputs with noises. Although the ANP model has been shown to work for image reconstruction tasks [32], it may fail for dynamics prediction in complex environments. In that case, one potential solution is to conduct dynamics prediction in the latent space as in Dreamer [44], which is directly applicable for CASRL. The hyperparameter-tuning for learning rates, network structures, and especially the latent dimensions could be another challenge for CASRL.

REFERENCES

[1] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, *et al.*, "Mastering the game of go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[2] O. Vinyals, I. Babuschkin, J. Chung, M. Mathieu, M. Jaderberg, W. M. Czarnecki, A. Dudzik, A. Huang, P. Georgiev, R. Powell,

et al., "Alphastar: Mastering the real-time strategy game starcraft ii," *DeepMind blog*, p. 2, 2019.

[3] A. Nagabandi, C. Finn, and S. Levine, "Deep online learning via meta-learning: Continual adaptation for model-based rl," *arXiv preprint arXiv:1812.07671*, 2018.

[4] B. Chen, M. Xu, L. Li, and D. Zhao, "Delay-aware model-based reinforcement learning for continuous control," *arXiv preprint arXiv:2005.05440*, 2020.

[5] J. Hwangbo, J. Lee, A. Dosovitskiy, D. Bellicoso, V. Tsounis, V. Koltun, and M. Hutter, "Learning agile and dynamic motor skills for legged robots," *Science Robotics*, vol. 4, no. 26, 2019.

[6] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[7] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[8] K. Chua, R. Calandra, R. McAllister, and S. Levine, "Deep reinforcement learning in a handful of trials using probabilistic dynamics models," in *Advances in Neural Information Processing Systems*, 2018, pp. 4754–4765.

[9] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 23–30.

[10] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," *arXiv preprint arXiv:1705.10528*, 2017.

[11] T.-H. Pham, G. De Magistris, and R. Tachibana, "Optlayer-practical constrained optimization for deep reinforcement learning in the real world," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 6236–6243.

[12] Y. Chow, O. Nachum, E. Duenez-Guzman, and M. Ghavamzadeh, "A lyapunov-based approach to safe reinforcement learning," in *Advances in neural information processing systems*, 2018, pp. 8092–8101.

[13] G. Dalal, K. Dvijotham, M. Vecerik, T. Hester, C. Paduraru, and Y. Tassa, "Safe exploration in continuous action spaces," *arXiv preprint arXiv:1801.08757*, 2018.

[14] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," *arXiv preprint arXiv:1703.03400*, 2017.

[15] A. Nagabandi, I. Clavera, S. Liu, R. S. Fearing, P. Abbeel, S. Levine, and C. Finn, "Learning to adapt in dynamic, real-world environments through meta-reinforcement learning," *arXiv preprint arXiv:1803.11347*, 2018.

[16] M. Xu, W. Ding, J. Zhu, Z. Liu, B. Chen, and D. Zhao, "Task-agnostic online reinforcement learning with an infinite mixture of gaussian processes," *arXiv preprint arXiv:2006.11441*, 2020.

[17] J. García and F. Fernández, "A comprehensive survey on safe reinforcement learning," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 1437–1480, 2015.

[18] E. Altman, *Constrained Markov decision processes*. CRC Press, 1999, vol. 7.

[19] E. Altman, "Constrained markov decision processes with total cost criteria: Lagrangian approach and dual linear program," *Mathematical methods of operations research*, vol. 48, no. 3, pp. 387–417, 1998.

[20] Y. Chow, O. Nachum, A. Faust, E. Duenez-Guzman, and M. Ghavamzadeh, "Lyapunov-based safe policy optimization for continuous control," *arXiv preprint arXiv:1901.10031*, 2019.

[21] T. Koller, F. Berkenkamp, M. Turchetta, and A. Krause, "Learning-based model predictive control for safe exploration," in *2018 IEEE Conference on Decision and Control (CDC)*. IEEE, 2018, pp. 6059–6066.

[22] L. Hewing, J. Kabzan, and M. N. Zeilinger, "Cautious model predictive control using gaussian process regression," *IEEE Transactions on Control Systems Technology*, 2019.

[23] A. Wachi and Y. Sui, "Safe reinforcement learning in constrained markov decision processes," in *International Conference on Machine Learning*. PMLR, 2020, pp. 9797–9806.

[24] A. Nilim and L. Ghaoui, "Robust markov decision problems with uncertain transition matrices," *Advances in Neural Information Processing Systems*, 2003.

[25] L. Pinto, J. Davidson, R. Sukthankar, and A. Gupta, "Robust adversarial reinforcement learning," *arXiv preprint arXiv:1703.02702*, 2017.

[26] W. Ding, B. Chen, M. Xu, and D. Zhao, "Learning to collide: An adaptive safety-critical scenarios generating method," 2020.

[27] L. Rice, E. Wong, and J. Z. Kolter, "Overfitting in adversarially robust deep learning," *arXiv preprint arXiv:2002.11569*, 2020.

[28] Y. Duan, J. Schulman, X. Chen, P. L. Bartlett, I. Sutskever, and P. Abbeel, "Rl^2: Fast reinforcement learning via slow reinforcement learning," *arXiv preprint arXiv:1611.02779*, 2016.

[29] S. Sæmundsson, K. Hofmann, and M. P. Deisenroth, "Meta reinforcement learning with latent variable gaussian processes," *arXiv preprint arXiv:1803.07551*, 2018.

[30] W. C. Cheung, D. Simchi-Levi, and R. Zhu, "Non-stationary reinforcement learning: The blessing of (more) optimism," *Available at SSRN 3397818*, 2019.

[31] M. Garnelo, J. Schwarz, D. Rosenbaum, F. Viola, D. J. Rezende, S. Eslami, and Y. W. Teh, "Neural processes," *arXiv preprint arXiv:1807.01622*, 2018.

[32] H. Kim, A. Mnih, J. Schwarz, M. Garnelo, A. Eslami, D. Rosenbaum, O. Vinyals, and Y. W. Teh, "Attentive neural processes," *arXiv preprint arXiv:1901.05761*, 2019.

[33] S. Qin, J. Zhu, J. Qin, W. Wang, and D. Zhao, "Recurrent attentive neural process for sequential data," *arXiv preprint arXiv:1910.09323*, 2019.

[34] J. Gordon, W. P. Bruinsma, A. Y. Foong, J. Requeima, Y. Dubois, and R. E. Turner, "Convolutional conditional neural processes," *arXiv preprint arXiv:1910.13556*, 2019.

[35] J. R. Pati, "Modeling, identification and control of cart-pole system," Ph.D. dissertation, 2014.

[36] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[37] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," *arXiv preprint arXiv:1511.05952*, 2015.

[38] C. J. Ostafew, A. P. Schoellig, and T. D. Barfoot, "Robust constrained learning-based nmpc enabling reliable mobile robot path tracking," *The International Journal of Robotics Research*, vol. 35, no. 13, pp. 1547–1563, 2016.

[39] A. Wächter and L. T. Biegler, "On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming," *Mathematical programming*, vol. 106, no. 1, pp. 25–57, 2006.

[40] M. P. Deisenroth, D. Fox, and C. E. Rasmussen, "Gaussian processes for data-efficient learning in robotics and control," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 2, pp. 408–423, 2013.

[41] A. Tamar, Y. Glassner, and S. Mannor, "Optimizing the cvar via sampling," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015, pp. 2993–2999.

[42] Z. Erickson, V. Gangaram, A. Kapusta, C. K. Liu, and C. C. Kemp, "Assistive gym: A physics simulation framework for assistive robotics," *IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

[43] T.-Y. Yang, J. Rosca, K. Narasimhan, and P. J. Ramadge, "Projection-based constrained policy optimization." in *ICLR*, 2020.

[44] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, "Dream to control: Learning behaviors by latent imagination," *arXiv preprint arXiv:1912.01603*, 2019.