Automatic Exploration Process Adjustment for Safe Reinforcement Learning with Joint Chance Constraint Satisfaction

Yoshihiro Okawa * **Tomotake Sasaki** * **Hidenao Iwane** **,[1]

* *Artificial Intelligence Laboratory, FUJITSU LABORATORIES LTD.,*
4-1-1 Kamikodanaka, Nakahara-ku, Kawasaki, Kanagawa 211-8588,
Japan (e-mail: okawa.y@fujitsu.com, tomotake.sasaki@fujitsu.com)
** *Research & IT Division, Reading Skill Test, Inc., 2-1-4 Shinkawa,*
Chuo-ku, Tokyo 104-0033, Japan (email: iwane@rstest.co.jp)

Abstract: In reinforcement learning (RL) algorithms, exploratory control inputs are used during learning to acquire knowledge for decision making and control, while the true dynamics of a controlled object is unknown. However, this exploring property sometimes causes undesired situations by violating constraints regarding the state of the controlled object. In this paper, we propose an automatic exploration process adjustment method for safe RL in continuous state and action spaces utilizing a linear nominal model of the controlled object. Specifically, our proposed method automatically selects whether the exploratory input is used or not at each time depending on the state and its predicted value as well as adjusts the variance-covariance matrix used in the Gaussian policy for exploration. We also show that our exploration process adjustment method theoretically guarantees the satisfaction of the constraints with the pre-specified probability, that is, the satisfaction of a joint chance constraint at every time. Finally, we illustrate the validity and the effectiveness of our method through numerical simulation.

Keywords: Reinforcement learning, Learning algorithm, Safe exploration, Safety-critical, Chance constraint

1. INTRODUCTION

Reinforcement learning (RL) methods are used to acquire knowledge for decision making and control, *i.e.*, "policy" or "control law" in online data-driven manners. In order to learn appropriate policies with RL methods, it is necessary to use exploratory control inputs, and they sometimes result in undesired situations. This is not a serious issue in the problems in which situations can be reset or recovered easily such as the video games (Mnih et al., 2015) or strategy board games (Silver et al., 2018); however, we cannot ignore this in many kinds of engineering problems. For addressing this issue, RL methods guaranteeing the "safety" during learning are demanded.

RL methods guaranteeing the "safety" have been discussed in literatures with different definitions of safety and some of them are called "safe reinforcement learning (safe RL) methods". According to the survey paper by García and Fernández (2015), approaches for safe RL are classified into two fundamental categories; "transforming the optimization criterion" and "modifying the exploration process". This paper adopts the latter one. In particular, we deal with the safety based on the constraints which are explicitly defined in control problems. More specifically, we define the probability of satisfaction of the constraints regarding the state of the controlled object as the quantitative index to evaluate the safety. In this paper, we propose an automatic exploration process adjustment method for

a Gaussian policy guaranteeing that the probability noted above is equal to or greater than its pre-specified lower bound, which is a kind of chance constraint (Shapiro et al., 2014). Our method automatically selects whether the exploratory input is used or not at each time depending on the state and its predicted value as well as adjusts the variance-covariance matrix of a normal distribution utilizing a linear nominal model of the controlled object and upper bounds of its approximation error. We show that an RL algorithm combined with our exploration process adjustment method theoretically guarantees the satisfaction of the joint chance constraint with the pre-specified probability.

The rest of this paper is organized as follows. In Section 2, we describe a problem formulation of this paper. Subsequently, in Section 3, we introduce our automatic exploration process adjustment method for the Gaussian policy and show a theoretical guarantee of probabilistic satisfaction of the constraints. In this section, we also show one concrete example of safe RL algorithms with our exploration process adjustment method. In addition, we compare our study with some related work for safe RL. In Section 4, we verify the validity and effectiveness of our method through numerical simulations, and finally, we conclude this paper in Section 5.

2. PROBLEM FORMULATION

Consider a discrete-time affine nonlinear system given by

---

[1] This work is done while he worked at Fujitsu Laboratories Ltd.

$$x_{k+1} = f(x_k) + g(x_k)u_k, \tag{1}$$

where $x_k \in \mathbb{R}^n$ and $u_k \in \mathbb{R}^m$ are the state and the control input at time k, respectively, and $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ and $g : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times m}$ are unknown nonlinear functions. We assume that the state x_k can be observed directly, and also, an instantaneous cost $c_k \in \mathbb{R}$ given by

$$c_{k+1} = \ell(x_k, u_k) \tag{2}$$

as well. Here $\ell : \mathbb{R}^n \times \mathbb{R}^m \rightarrow [0, \infty)$ is an unknown instantaneous cost function. As described in the left hand side of (2), we denote the instantaneous cost by c_{k+1} following the convention adopted in Sutton and Barto (2018), while it is produced with the state x_k and the input u_k at time k.

The objective of control is to minimize the cumulative discounted cost J given by

$$J = \sum_{k=0}^{\infty} \gamma^k c_{k+1}, \tag{3}$$

where $\gamma \in (0, 1]$ is a discount rate.

In addition, we consider the situation where desirable conditions about the state is given by the following linear inequalities:

$$Hx \preceq v, \tag{4}$$

where $v = [v_1, \ldots, v_{n_c}]^{\mathrm{T}} \in \mathbb{R}^{n_c}$ and $H = [h_1, \ldots, h_{n_c}]^{\mathrm{T}} \in \mathbb{R}^{n_c \times n}$, and the symbol $\preceq$ represents that every inequality $\leq$ on $\mathbb{R}$ is satisfied regarding each component in the vector. We suppose that the inequalities (4) is known, and we define the safety in this paper based on them.

Let the set $\mathcal{X} \subset \mathbb{R}^n$ be

$$\mathcal{X} := \{x \in \mathbb{R}^n \mid Hx \preceq v\}, \tag{5}$$

and $\mathcal{X}^{\mathrm{int}}$ be the interior point set of $\mathcal{X}$. We suppose the existence of the state $x^* \in \mathcal{X}^{\mathrm{int}}$ which satisfies $f(x^*) = x^*$ and $\ell(x^*, 0) = 0$, while we do not suppose that x^* is known. This means that some kind of consistency is held between the cost and safety in the control problem, and the state satisfying both of them is maintainable. For simplicity, we also suppose that the initial state x_0 satisfies $x_0 \in \mathcal{X}$.

Next, we describe the goal of this paper. The common goal of RL methods is to acquire a policy, which corresponds to a control law in control engineering, to minimize (or maximize) evaluation functions related to the instantaneous cost (or the instantaneous reward) obtained at each time. If we know both complete dynamics of the controlled object and the evaluation function, the problem leads to the usual optimal control problem. However, in many cases the above-mentioned information is not completely known beforehand, and the usual optimal control methods cannot attain desired performance. A fundamental appeal of RL methods is that we can obtain the optimal policy or at least an improved one even in such a situation by updating the current policy with the information obtained online.

In order to acquire an appropriate policy while the properties of the controlled object are unknown, we have to select exploratory inputs during learning which are not optimal according to the current policy. This means that, the possibility of violating desirable conditions given by (4) increases because of the exploration.

In general, stochastic policies are used in RL methods to carry out the above exploration, and thus, the state x_k of the controlled object also transits stochastically. To evaluate the satisfaction of the conditions given by (4) quantitatively, we consider the (joint) chance constraint

$$\Pr\{Hx_k \preceq v\} \geq \eta. \tag{6}$$
$$(\Leftrightarrow \Pr\{h_j^{\mathrm{T}} x_k \preceq v_j, \forall j = 1, 2, \cdots, n_c\} \geq \eta)$$

In the above inequality, $\Pr\{\cdot\}$ denotes the probability of the satisfaction of the inequalities in $\{\cdot\}$. We call this "probability of constraint satisfaction" in the rest of this paper, and use it as the metric of safety. For simplicity, we assume $\eta \in (0.5, 1)$.

The goal of this paper is to propose an automatic exploration process adjustment method for safe RL which guarantees the satisfaction of the joint chance constraint (6) at every time $k \geq 1$ with η determined before learning. [2]

Throughout this paper, we assume that the following two most basic conditions are satisfied.

Assumption 1. The following linear approximate (nominal) model of the nonlinear system in (1) is known:

$$x_{k+1} \simeq Ax_k + Bu_k, \tag{7}$$

where $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times m}$.

Let $e(\cdot, \cdot; f, g, A, B) : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$ be the following function which denotes the approximation error between the nonlinear system given in (1) and its linear approximate model given in (7):

$$e(x, u; f, g, A, B) := f(x) + g(x)u - (Ax + Bu) \tag{8}$$
$$=: [e_1(x, u; f, g, A, B), \ldots, e_n(x, u; f, g, A, B)]^{\mathrm{T}}. \tag{9}$$

We make the next assumption regarding this function.

Assumption 2. Regarding the functions $e_i(\cdot, \cdot; f, g, A, B)$, $i = 1, 2, \ldots, n$ defined in (9), scalar values $\bar{e}_i \in [0, \infty)$, $i = 1, 2, \ldots, n$ satisfying the following relation are known:

$$\bar{e}_i \geq \sup_{x \in \mathbb{R}^n, u \in \mathbb{R}^m} |e_i(x, u; f, g, A, B)|. \tag{10}$$

An example satisfying this assumption is given in Section 4.1.

Under Assumption 2, let $\mathcal{E} \subset \mathbb{R}^n$ be the set of the vectors $\epsilon = [\epsilon_1, \ldots, \epsilon_n]^{\mathrm{T}} \in \mathbb{R}^n$ whose elements are

$$\epsilon_i = \bar{e}_i \text{ or } -\bar{e}_i. \tag{11}$$

Note that the size of the set $\mathcal{E}$ is 2^n by definition.

Assumptions 1 and 2 mean that we do not know the exact system dynamics but have certain level of prior knowledge about it somehow (e.g. by physical considerations). We introduce three other assumptions right before we use them in the next section.

We describe our proposed method in detail and show a theorem which theoretically guarantees the joint chance constraint satisfaction in the following section.

Remark 1. In order to simplify the notations in the rest of this paper, we generalize the joint chance constraint (6) and denote $\Pr\{H\xi \preceq v\} \geq \lambda$ for an arbitrary $\xi \in \mathbb{R}^n$ and $\lambda \in (0.5, 1)$ by $\mathsf{CC}(\xi, \lambda)$. That is,

$$\mathsf{CC}(\xi, \lambda) \overset{\text{def}}{\Longleftrightarrow} \Pr\{H\xi \preceq v\} \geq \lambda. \tag{12}$$

---

[2] In this paper, we describe the chance constraint (6) as a joint chance constraint, while it is not a "joint" one if $n_c = 1$.

For example, with this notation, the joint chance constraint (6) can be denoted by $\mathsf{CC}(x_k, \eta)$.

In addition, it is sometimes difficult to deal with the joint chance constraint (6) directly since it is a condition with respect to the probability of the simultaneous satisfaction of all n_c constraints. Therefore, we show some supplemental theoretical results with respect to individual constraints before showing our main result in Theorem 1. For this purpose, similar to the notation $\mathsf{CC}(\cdot, \cdot)$, we denote $\left\{ \Pr\left\{ h_j^{\mathrm{T}} \xi \leq v_j \right\} \geq \lambda, \forall j = 1, \ldots, n_c \right\}$ for an arbitrary $\xi \in \mathbb{R}^n$ and $\lambda \in (0.5, 1)$ by $\mathsf{CC}'(\xi, \lambda)$. That is,

$$\mathsf{CC}'(\xi, \lambda) \overset{\text{def}}{\iff} \left\{ \Pr\left\{ h_j^{\mathrm{T}} \xi \leq v_j \right\} \geq \lambda, \forall j = 1, \ldots, n_c \right\}. \quad (13)$$

Regarding the relationship between the above two kinds of chance constraints (12) and (13), we can easily prove based on Bonferroni's inequality that $\mathsf{CC}'(\xi, 1 - \frac{1-\lambda}{n_c})$ is a sufficient condition for $\mathsf{CC}(\xi, \lambda)$.

3. MAIN RESULT

As described in Section 1, approaches for safe RL are classified into two fundamental categories; "transforming the optimization criterion" and "modifying the exploration process". We adopt the latter one. In the rest of this section, Subsection 3.1 shows the case in which the exploration is completely removed in the RL algorithm as preliminaries. Next, based on the results shown in Subsection 3.1, we introduce an automated exploration process adjustment method for the Gaussian policy and show a theorem which gives our method a theoretical guarantee regarding the satisfaction of the joint chance constraint in Subsection 3.2. Finally, in Subsection 3.3, we show an example of safe RL methods based on our exploration process adjustment method. We compare our study with related work for safe RL in Section 3.4.

3.1 Chance constraint satisfaction regarding the input without exploration

When we consider satisfaction of the constraints by modifying the exploration process in an RL algorithm, it is natural to use an input without exploration in some situations. However, even if we do so, the satisfaction of the chance constraint is not always guaranteed. Therefore, we have to select the input carefully even in such a situation. In this subsection, we show two particular cases in which the input can be selected to guarantee the satisfaction of the chance constraint.

Firstly, let us consider the case in which the function f in (1) obeys the following assumption.

Assumption 3. If $x \in \mathcal{X}$ holds, then $f(x) \in \mathcal{X}$ holds.

It is straightforward to see $\Pr\{Hx_{k+1} \preceq v\} = 1$ holds if we use $u_k = 0$ for $x_k \in \mathcal{X}$ in this case.

Next, we consider the case in which the particular inputs to return the state into $\mathcal{X}$ from the outside of $\mathcal{X}$ can be obtained. Specifically, we make the following assumption.

Assumption 4. Suppose that $x_k = x \notin \mathcal{X}$ at time $k \geq 1$. Regardless of time k and the state x, we can make the state be in $\mathcal{X}$ within τ step with the particular successive inputs. That is, we previously know the successive inputs

$u_k^{back}, u_{k+1}^{back}, \ldots, u_{k+j-1}^{back}$ or its calculation procedure to let the state be $x_{k+j} \in \mathcal{X}$ $(1 \leq j \leq \tau)$ for any k and x.

In this case, we obtain the following lemma.

Lemma 1. Let Assumption 4 hold. In addition, we assume that $\mathsf{CC}(x_{k+1}, p)$ is satisfied if $x_k \in \mathcal{X}$. Then, $\mathsf{CC}(x_{k+1}, p^{\tau})$ is satisfied if $x_0 \in \mathcal{X}$.

Proof. At time k, we define "$x_k \in \mathcal{X}$" as State 1 and "$x_{k-i} \in \mathcal{X} \wedge (x_{k-i+1}, \ldots, x_k \notin \mathcal{X})$" as State $i+1$. With this definition, the state transition is expressed by the Markov chain in Proposition 1 described in Appendix A, and thus, this lemma is proven with $x_0 \in \mathcal{X}$ and $\rho_1 \geq p$. $\quad\square$

3.2 Automatic exploration process adjustment method for Gaussian policy

Next, we propose an automatic exploration process adjustment method which includes usage of the inputs with exploration. In this method, we basically generate the inputs based on a Gaussian policy, which can be applied to problems in continuous state and action space. Specifically, we use the following Gaussian probability density function as the policy function:

$$\Pi(u \mid x; \ w, \Sigma)$$
$$= \frac{1}{(\sqrt{2\pi})^m \sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(u - \mu(x; w))^{\mathrm{T}} \Sigma^{-1} \bullet \right), \quad (14)$$

where $\bullet$ represents the omission of $(u - \mu(x; w))$ and $w \in \mathbb{R}^{N_w}$ is the policy parameter. We generate an input u stochastically according to the m-dimensional normal distribution with the mean $\mu(x; w) \in \mathbb{R}^m$ and the variance-covariance matrix $\Sigma \in \mathbb{R}^{m \times m}$. We express this as

$$u \sim \mathcal{N}\left(\mu(x; w), \Sigma \right). \quad (15)$$

The degree of exploration, in other words, how different input is selected from its mean, depends on Σ.

Let μ_k denote the mean of the input corresponding to the state x_k of the controlled object at time k and the policy parameter w_k, that is,

$$\mu_k := \mu(x_k; w_k). \quad (16)$$

Now we restrict the variance-covariance matrix Σ to the following diagonal matrix determined by σ^2 $(\sigma > 0)$:

$$\Sigma = \sigma^2 I_m, \quad (17)$$

where I_m is an $m \times m$ identity matrix.

Consider the case in which the following assumption holds between the linear nominal model of the controlled object and the constraints.

Assumption 5. The coefficient matrix B of the nominal model in (7) and $H = [h_1, \ldots, h_{n_c}]^{\mathrm{T}}$ in the constraints in (4) satisfies

$$h_j^{\mathrm{T}} B \neq 0, \quad \forall j = 1, 2, \ldots, n_c. \quad (18)$$

In this case, we have the following lemma regarding the relationship between the standard deviation σ of the Gaussian policy and the satisfaction of the joint chance constraint.

Lemma 2. Let Assumptions 1, 2 and 5 hold, and let the input u_k be selected according to the normal distribution
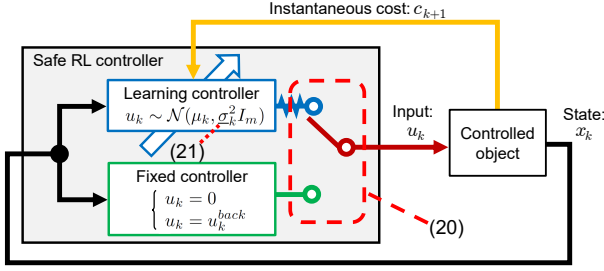
Fig. 1. Block diagram of the closed loop system with our automatic exploration process adjustment method

with mean μ_k and variance-covariance matrix $\sigma^2 I_m$. In addition, assume $h_j^{\mathrm{T}}(Ax_k + B\mu_k + \epsilon) < v_j, \forall j = 1, 2, \ldots, n_c$. Then,

$$\sigma = \min_{j,\epsilon} \frac{1}{\|h_j^{\mathrm{T}}B\|_2 \Phi^{-1}(q)} \left\{ v_j - h_j^{\mathrm{T}}(Ax_k + B\mu_k + \epsilon) \right\} \quad (19)$$

is a sufficient condition for $\mathsf{CC}'(x_{k+1}, q)$ being held with $q \in (0.5, 1)$. Here $\Phi(\cdot)$ is the cumulative distribution function of a zero mean unit variance Gaussian random variable.

Proof. The conclusion follows from Corollary 3 in Appendix B and the assumption $h_j^{\mathrm{T}}(Ax_k + B\mu_k + \epsilon) < v_j$, $\forall j = 1, 2, \ldots, n_c$. □

In the rest of this paper, we assume that Assumptions 1-5 hold. Based on the results described in Subsection 3.1 and Lemma 2, we propose the following control law as an automatic exploration process adjustment method for the Gaussian policy.

$$\begin{cases} u_k \sim \mathcal{N}(\mu_k, \underline{\sigma}_k^2 I_m) & \text{if } x_k \in \mathcal{X} \wedge (\bar{x}_{k+1}^\epsilon \in \mathcal{X}^{\mathrm{int}}, \forall \epsilon \in \mathcal{E}) \\ u_k = 0 & \text{else if } x_k \in \mathcal{X} \\ u_k = u_k^{back} & \text{otherwise} \end{cases}, \quad (20)$$

where $\bar{x}_{k+1}^\epsilon := Ax_k + B\mu_k + \epsilon$. In addition, we define $\underline{\sigma}_k$ in the top of (20) as

$$\underline{\sigma}_k := \min_{j,\epsilon} \frac{1}{\|h_j^{\mathrm{T}}B\|_2 \Phi^{-1}(\eta')} \left(v_j - h_j^{\mathrm{T}} \bar{x}_{k+1}^\epsilon \right), \quad (21)$$

where $\eta' := 1 - \frac{1 - \eta^{\frac{1}{\tau}}}{n_c}$. The block diagram of the closed loop system including the safe RL controller based on our exploration process adjustment method is depicted in Fig. 1. As shown in this figure, the safe RL controller selects whether the exploratory input is used or not at each time according to the control law (20) as well as adjusts the variance-covariance matrix of a normal distribution used in the Gaussian policy for exploration with $\underline{\sigma}_k$ in (21). In other words, the safe RL controller contains two kinds of controllers; the learning controller with a time-varying variance-covariance matrix and the model-based fixed one, and selects either of them to generate the control input at each time[3].

We have the following theorem regarding the satisfaction of the joint chance constraint with the above control law as an automatic exploration process adjustment method.

Theorem 1. Let Assumptions 1-5 hold and the input u_k at time k be selected according to (20). Then, the joint

---

[3] According to the terminology introduced in Okawa et al. (2019), this is classified as "switching LeFiCo".

Algorithm 1 Safe one-step actor-critic algorithm with automatic exploration process adjustment

1: **Initialize:**
2:    $\theta \leftarrow \theta_0, \ w \leftarrow w_0, \ \iota \leftarrow 1$
3: **loop** (Execute below at $t = 0$ and every T_s time)
4:    Observe x
5:    **if** $t \neq 0$ **then**
6:       Get cost c
7:       $\delta \leftarrow -c + \gamma \hat{V}(x; \theta) - \hat{V}(x^-; \theta)$
8:       $\theta \leftarrow \theta + \alpha \delta \frac{\partial \hat{V}}{\partial \theta}(x^-; \theta)$
9:       $w \leftarrow w + \beta \iota \delta \frac{\partial \log \Pi}{\partial w}(u^- | x^-; w)$
10:   **end if**
11:   Select u according to (20)
12:   Input u to controlled object
13:   $x^- \leftarrow x, \ u^- \leftarrow u, \ \iota \leftarrow \gamma \iota$
14: **end loop**

---

chance constraint (6), that is, $\mathsf{CC}(x_k, \eta)$, is satisfied at all time $k \geq 1$.

Proof. If $x_k \in \mathcal{X} \wedge (\bar{x}_{k+1}^\epsilon \in \mathcal{X}^{\mathrm{int}}, \forall \epsilon \in \mathcal{E})$, $\mathsf{CC}'(x_{k+1}, \eta')$ is satisfied by selecting the inputs stochastically according to $u_k \sim \mathcal{N}(\mu_k, \underline{\sigma}_k^2 I_m)$ from Lemma 2, and this is a sufficient condition for $\mathsf{CC}(x_{k+1}, \eta^{\frac{1}{\tau}})$. In addition, if $x_k \in \mathcal{X} \wedge \neg(\bar{x}_{k+1}^\epsilon \in \mathcal{X}^{\mathrm{int}}, \forall \epsilon \in \mathcal{E})$, x_{k+1} always satisfies $x_{k+1} \in \mathcal{X}$ with the input $u_k = 0$ from Assumption 3. Therefore, $\mathsf{CC}(x_{k+1}, \eta^{\frac{1}{\tau}})$ is satisfied at all time $k \geq 1$ with the control law (20) if $x_k \in \mathcal{X}$. Furthermore, this $\mathsf{CC}(x_{k+1}, \eta^{\frac{1}{\tau}})$ is a sufficient condition for $\mathsf{CC}(x_{k+1}, \eta)$ since $\eta^{\frac{1}{\tau}} \geq \eta$ for any $\tau \in \mathbb{Z}_{\geq 1}$.

On the other hand, if $x_k \notin \mathcal{X}$, $k \geq 1$, Lemma 1 shows that $\mathsf{CC}(x_{k+1}, \eta)$ is satisfied for an arbitrary $x_k \in \mathbb{R}^n$ with the input $u_k = u_k^{back}$ according (20).

Consequently, $\mathsf{CC}(x_k, \eta)$ is satisfied at all time $k \geq 1$. □

Remark 2. As shown in the proof described in Appendix B, the pair (B.7) and (B.8) is a sufficient condition for the state x_{k+1} at the next time $k + 1$ to satisfy the joint chance constraint (6) for more general Σ. In this case, however, it is difficult to derive Σ which satisfies this condition.

Remark 3. Theorem 1 requires that the controlled object satisfies Assumption 3. However, there may not be many systems satisfying this assumption by themselves. For this problem, we can expand the applicability of our proposed method by forming a minor feedback loop with some kind of pre-designed controllers based on known information of the controlled object. We can say that this is one of the advantages of the control approach to use RL and model-based control simultaneously in parallel ("parallel LeFiCo") as discussed in Okawa et al. (2019).

3.3 Safe learning algorithm with the automatic exploration process adjustment

Algorithm 1 shows a concrete example of the proposed safe RL method, which is the algorithm of the one-step actor-critic described in the Section 13 of Sutton and Barto (2018) with our automatic exploration process adjustment method shown in the previous subsection. In this algorithm, $\hat{V}(x; \theta)$ is the estimated value of the state

value function, $\theta \in \mathbb{R}^{N_\theta}$ is the sate value weight, and N_θ is its number. In addition, $\delta \in \mathbb{R}$ is the TD (temporal difference) error, and $\alpha \in [0,1)$ and $\beta \in [0,1)$ are the learning rates (step sizes). In this algorithm, the inputs are selected according to the control law (20) to adjust its exploration process automatically at each time as shown in Line 11, and then the state value weight θ and the policy parameter w are updated. As a result, we achieve to learn the appropriate policy by updating policy parameters with the satisfaction of the joint chance constraint (6) for the pre-specified η guaranteeing that the probability of constraint satisfaction is equal to or greater than its pre-specified lower bound.

3.4 Comparison with related work

In the field of RL, the concept of "safety" (or its opponent, "risk") is defined according to many kinds of formulation. Here we restrict our discussions to the ones regarding continuous state and action spaces as ours. For example, Wen and Topcu (2018) proposed an RL method in which constraints satisfied for safety are defined as the expected cost over finite-length trajectories. In addition, some RL methods have been proposed to guarantee their safety even during learning. Berkenkamp et al. (2017) used a known policy which guarantees safety to prevent the system from getting into an unrecoverable or undesired situation. Also, Achiam et al. (2017) proposed a policy search algorithm for an RL problem in a constrained Markov Decision Process, which guarantees constraint satisfaction throughout training. However, these studies do not deal with satisfaction of the constraints explicitly defined in their control problems as their safety.

The following three methods regarding safe RL are the most similar to our study in the sense that they guarantee their safety during learning from viewpoint of satisfying the constraints explicitly defined in their control problems. Dalal et al. (2018) showed a safe exploration method for RL algorithms to satisfy its safety constraints if one-time initial pre-training of a model can be used. Li et al. (2018) proposed a safe RL framework with a supervisory element between the RL agent and the linear control system. Furthermore, Cheng et al. (2019) showed how to modify existing RL algorithms to guarantee safety of the nonlinear system whose dynamics consists of partially known autonomous dynamics and completely known actuated one. However, as compared with these existing methods, the safe RL method with our exploration process adjustment method guarantees its safety theoretically even if we can only use partial information of both autonomous and actuated dynamics of the nonlinear system: a linear approximate (nominal) model and upper bounds of its approximation error.

4. SIMULATION VERIFICATION

This section verifies the validity and effectiveness of the safe RL method with our automatic exploration process adjustment method through numerical simulation.

4.1 Simulation condition

Control objective and constraints   Let us consider the following nonlinear function $f : \mathbb{R}^2 \to \mathbb{R}^2$:

$$f(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix} = \begin{bmatrix} 0.3x_1 - 0.4\sin x_2 \\ -0.1x_2 + 0.2\cos x_1 - 0.2 \end{bmatrix}, \qquad (22)$$

where $x = [x_1, x_2]^\mathrm{T} \in \mathbb{R}^2$ and $f(0) = 0$. Regarding this function, the Jacobian matrix is given by

$$\frac{\partial f}{\partial x}(x) = \begin{bmatrix} 0.3 & -0.4\cos x_2 \\ -0.2\sin x_1 & -0.1 \end{bmatrix}.$$

Therefore, the Frobenius norm $\left\| \frac{\partial f}{\partial x}(x) \right\|_F$ of $\frac{\partial f}{\partial x}(x)$ satisfies

$$\left\| \frac{\partial f}{\partial x}(x) \right\|_F = \sqrt{0.04\sin^2 x_1 + 0.16\cos^2 x_2 + 0.1} < 1$$

for an arbitrary $x \in \mathbb{R}^2$ since $|\sin x_1| \le 1$ and $|\cos x_2| \le 1$. This means that f is a contraction mapping in entire $\mathbb{R}^2$ with the origin being its fixed point.

We let $x_k = [x_{1_k}, x_{2_k}]^\mathrm{T} \in \mathbb{R}^2$ and $u_k \in \mathbb{R}$ be the state and the input at time k, respectively, and consider the following discrete-time affine nonlinear system:

$$x_{k+1} = f(x_k) + gu_k, \qquad (23)$$

where f is the function described above and $g = [1,1]^\mathrm{T}$. In addition, we use the following linear nominal model:

$$x_{k+1} \simeq Ax_k + bu_k, \quad A = \begin{bmatrix} 0.3 & 0 \\ 0 & -0.1 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \qquad (24)$$

According to Assumption 1, we assume A and b are known while f and g are unknown. The upper bounds of the error function e_i given in (9) become

$$\sup_{x \in \mathbb{R}^2, u \in \mathbb{R}} |-0.4\sin x_2| = 0.4, \quad \sup_{x \in \mathbb{R}^2, u \in \mathbb{R}} |0.2\cos x_1 - 0.2| = 0.4.$$

Therefore, we let $\bar{e}_1 = 0.4$ and $\bar{e}_2 = 0.4$, and assume that these values are previously known to satisfy Assumption 2.

On the other hand, we let $|x_1| \le 10$ be the constraints. That is, the set $\mathcal{X}$ is given by $\mathcal{X} = \{ x \in \mathbb{R}^2 \mid Hx \preceq v \}$, where

$$H = \begin{bmatrix} h_1^\mathrm{T} \\ h_2^\mathrm{T} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1 & 0 \end{bmatrix}, \quad v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 10 \\ 10 \end{bmatrix}.$$

Assumption 3 is satisfied since f is a contraction mapping in $\mathbb{R}^2$ with the origin being its fixed point as described above and $\mathcal{X}$ is a convex region containing the origin. We also let the initial state be $x_0 = [5,5]^\mathrm{T} \in \mathcal{X}$.

Since $h_j^\mathrm{T} b \ne 0$, $j \in \{1,2\}$, the pair of the above nominal model and the constraints satisfies Assumption 5.

In addition, we assume that the instantaneous cost $c_{k+1} = \ell(x_k, u_k)$ can be measured directly at each time, while the following cost function $\ell(\cdot, \cdot)$ is unknown:

$$\begin{aligned} \ell(x_k, u_k) &= x_{k+1}^\mathrm{T} Q x_{k+1} + R u_k^2 \\ &= (f(x_k) + gu_k)^\mathrm{T} Q(f(x_k) + gu_k) + R u_k^2, \quad (25) \end{aligned}$$

where $Q = 1.0 \times 10^5 I_2$ and $R = 1$.

Control input determination with the proposed exploration process adjustment method   In this verification, we update a policy according to Algorithm 1 described in Subsection 3.3. Specifically, we define $T(= 15)$ steps as 1 episode and learn the policy to determine inputs which minimize the cumulative cost $J = \sum_{k=0}^{T-1} c_{k+1}$ of the instantaneous cost given in (25) in each episode starting from the initial state x_0.

The estimated state value function $\hat{V}(x;\theta)$ and the mean $\mu(x;w)$ of the input u are, respectively, given by

Table 1. Simulation parameters

Symbol	Definition	Value
T	Number of simulation steps	15
γ	Discount rate	1.0
α, β	Learning rates	1.0×10^{-10}
N_θ, N_w	Number of learning parameters	121
N	Number of learning episodes	1.5×10^4
η	Lower bound of probability of constraint satisfaction	0.95

$$\hat{V}(x;\theta) = \sum_{i=1}^{N_\theta} \phi_i(x)\theta_i, \ \mu(x;w) = \sum_{i=1}^{N_w} \phi_i(x)w_i, \quad (26)$$

where $\theta = [\theta_1, \ldots, \theta_{N_\theta}]^{\mathrm{T}} \in \mathbb{R}^{N_\theta}$ and $w = [w_1, \ldots, w_{N_w}]^{\mathrm{T}} \in \mathbb{R}^{N_w}$. In the above equation, $\phi_i : \mathbb{R}^2 \to \mathbb{R}$ represent the feature extractors given by the following Gaussian radial basis functions:

$$\phi_i(x) = \exp\left(-\frac{\|x - \varrho_i\|^2}{2\varsigma_i^2}\right), \ i = 1, \ldots, N_\theta, \quad (27)$$

where $\varrho_i \in \mathbb{R}^2$ and $\varsigma_i^2 > 0$ are the central points and the variances of each basis function, respectively. We calculate the mean value $\mu_k = \mu(x_k; w_k)$ according to (26) with the state x_k at each time and the policy parameters w_k, and then, select the input u_k according to the control law (20). On the other hand, the pair of the coefficient matrix (A, b) of the nominal model in (24) is controllable and its controllability index is 2. From this result, we calculate the inputs u_k^{back} in the control law (20) as follows:

$$[u_k^{back}, u_{k+1}^{back}]^{\mathrm{T}} = \tilde{B}^{-1}(x_0 - A^2 x_k), \ \tilde{B} := [Ab, b]. \quad (28)$$

These inputs are designed so as to make the state go back to the initial point. Indeed, these two successive inputs make any state $x \notin \mathcal{X}$ return into $\mathcal{X}$ within 2 steps. Therefore, Assumption 4 is satisfied. The state value weight θ and the policy parameter w are updated with the instantaneous cost given by (25) at each time. The values of each parameter are listed in Table 1.

4.2 Simulation result

The results of this verification are shown in Figs. 2-7. Fig. 2 shows the cumulative cost J obtained without exploration, that is, letting the inputs u_k at each step time be their mean values μ_k calculated with the policy parameters w_k, at every 50 episodes. Fig. 3 shows the relative frequencies of the constraint satisfaction with respect to each time which are obtained by dividing the number of the episodes satisfying the constraints by its total number N, and Fig. 4 is the enlarged figure of Fig. 3. Since the total number of episodes N is large enough and the occurrence of constraint satisfaction between arbitrary two episodes is independent of each other, we use relative frequencies to confirm the validity of Theorem 1. In addition, Fig. 5 shows the margins from the constraints and its minimum value at the final episode. Due to the simulation conditions described in the previous subsection, the margins $v_j - h_j^{\mathrm{T}}(Ax_k + B\mu_k + \epsilon), \epsilon \in \mathcal{E}, j \in \{1, 2\}$ can be summarized into four scaler values. We denote them by $\Delta_k(i, j)$ where $i = 1$ if $\epsilon = [\bar{e}_1, \pm\bar{e}_2]^{\mathrm{T}}$ and $i = 2$ if $\epsilon = [-\bar{e}_1, \pm\bar{e}_2]^{\mathrm{T}}$, and show $\Delta_k(i, j)$ in dashed lines and its minimum value at each step time in a solid line in this figure, respectively. Fig. 6 shows $\underline{\sigma}_k$ given in (21) at the final episode. Furthermore, Fig. 7 shows the trajectory of
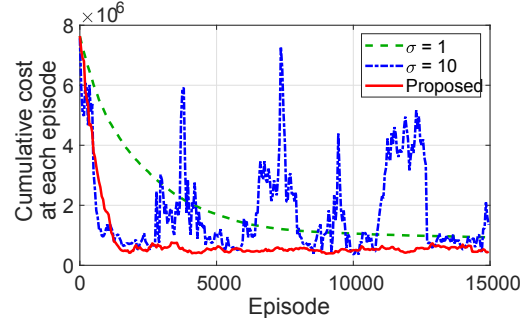

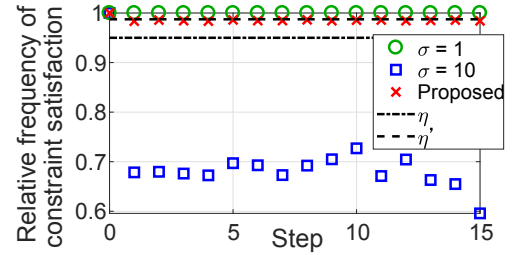Fig. 2. Cumulative cost at each episode

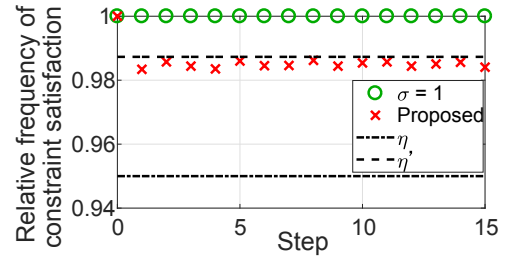Fig. 3. Relative frequencies of constraint satisfaction

Fig. 4. Enlarged figure of Fig. 3

the state with the policy parameter at the final episode. Among these figures, in Figs. 2-4 and 7, the results by using the Gaussian policy with the fixed standard deviation $\sigma = 1$ are shown in green, those with the bigger fixed standard deviation $\sigma = 10$ are in blue, and those by using our proposed method are in red. Furthermore, in this verification, if $|x_{1_k}| > 10, k = 1, 2, \ldots, T$ at each episode, we gave a penalty $(T - k + 1)c_{k+1}$ and let the estimated state value be $\hat{V}(x) = 0$ to terminate the current episode, and then, start another episode from the initial state when we use the Gaussian policy with fixed standard deviations.

We can confirm that the relative frequencies of the constraint satisfaction with the fixed standard deviation $\sigma = 1$ in green is greater than the lower bound of the probability of constraint satisfaction $\eta = 0.95$ in Fig. 3; however, as shown in Fig. 2, the corresponding result of the cumulative cost decreases slowly as the number of episodes increases. On the other hand, the results with the fixed standard deviation $\sigma = 10$ in blue show that, though the cumulative cost decreases efficiently in Fig. 2, there are some step times when the relative frequency of the constraint satisfaction becomes lower than η. As compared with these results, our proposed method achieves not only to decease its cumulative cost efficiently but also to guarantee that the relative frequencies of the constraint satisfaction are greater than $\eta = 0.95$ at all step times, while there are some step times when its value become lower than
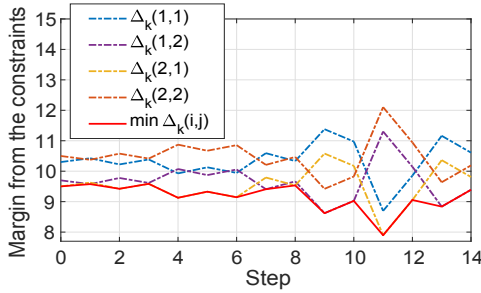
Fig. 5. Results of margins from the constraints and its minimum at the final episode
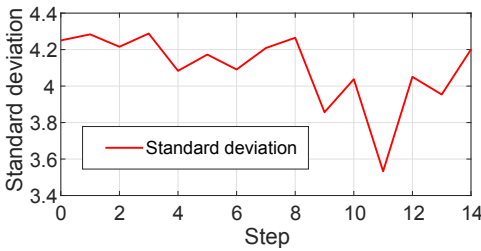
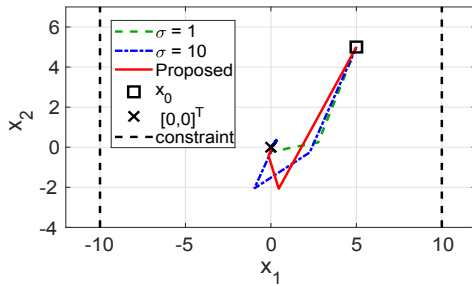

Fig. 6. Result of $\underline{\sigma}_k$ at the final episode

Fig. 7. Result of the trajectory

$\eta' = 1 - \frac{1-\eta^{\frac{1}{2}}}{2} \approx 0.987$ which is used to derive the standard deviation $\underline{\sigma}_k$ in (21) as shown in Fig. 4.

In addition, it is shown that the standard deviation $\underline{\sigma}_k$ in Fig. 6 corresponds to the minimum value among $\Delta_k(i,j)$ at each step time in a solid line in Fig. 5. We can confirm from this result that our proposed method automatically adjusts its exploration process by changing the variance $\underline{\sigma}_k^2$ in the Gaussian policy at each step time, even if the input is used for exploration.

Furthermore, from the control result with the policy parameters at the final episode shown in Fig. 7, the learned policy parameter enables us to transit the state x_k of the system into the origin $x = [0,0]^{\mathrm{T}}$, which is the fixed point of $f(x)$ and the instantaneous cost given in (25) becomes 0 with the input $u_k = 0$.

5. CONCLUSION AND FUTURE WORK

In this paper, we considered an RL problem in continuous state and action spaces with constraints explicitly defined in the control problem, and proposed an automatic exploration process adjustment method for safe RL which achieves the satisfaction of a joint chance constraint derived from the above-mentioned constraints. More specif-

ically, our method adjusts the exploration process automatically utilizing a known linear nominal model of the controlled object, and we theoretically showed that this method attains the pre-specified lower bound of the constraint satisfaction probability. We also verified the validity and effectiveness of our method through numerical simulations and showed that it achieves to learn its policy appropriately by guaranteeing the satisfaction of the joint chance constraint with the pre-specified probability.

One of the future work is to develop a safe RL method for control problems in which reinforcement learning methods and model-based control methods are used simultaneously with different sampling times. We have to consider this situation since updating huge number of parameters in an RL method requires more computation in general, and thus, it is required to take a longer time to determine its control input. However, in such a situation, the probability of constraint satisfaction is not guaranteed with the exploration process adjustment method presented in this paper since the state transition with the control inputs from a model-based control method between the sampling times of an RL method is not considered. Therefore, we plan to develop an automatic exploration process adjustment method to guarantee the satisfaction of the constraints even in such a situation.

REFERENCES

Achiam, J., Held, D., Tamar, A., and Abbeel, P. (2017). Constrained policy optimization. In *the 34th International Conference on Machine Learning*, 22–31.

Berkenkamp, F., Turchetta, M., Schoellig, A., and Krause, A. (2017). Safe model-based reinforcement learning with stability guarantees. In *Advances in Neural Information Processing Systems*, 908–918.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Cheng, R., Orosz, G., Murray, R.M., and Burdick, J.W. (2019). End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In *Thirty-Third AAAI Conference on Artificial Intelligence*, 3387–3395.

Dalal, G., Dvijotham, K., Vecerík, M., Hester, T., Paduraru, C., and Tassa, Y. (2018). Safe exploration in continuous action spaces. *CoRR*, abs/1801.08757. URL http://arxiv.org/abs/1801.08757.

García, J. and Fernández, F. (2015). A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1), 1437–1480.

Li, Z., Kalabić, U., and Chu, T. (2018). Safe reinforcement learning: Learning with supervision using a constraint-admissible set. In *2018 Annual American Control Conference (ACC)*, 6390–6395. IEEE.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540), 529–533.

Okawa, Y., Sasaki, T., and Iwane, H. (2019). Control approach combining reinforcement learning and model-based control. In *2019 12th Asian Control Conference (ASCC)*, 1419–1424. IEEE.

Shapiro, A., Dentcheva, D., and Ruszczyński, A. (2014). *Lectures on Stochastic Programming: Modeling and Theory, Second Edition.* MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., and Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419), 1140–1144.

Sutton, R.S. and Barto, A.G. (2018). *Reinforcement learning: An introduction.* MIT Press, 2nd edition.

Wen, M. and Topcu, U. (2018). Constrained cross-entropy method for safe reinforcement learning. In *Advances in Neural Information Processing Systems*, 7450–7460.

Appendix A. PROPERTY OF MARKOV CHAIN

Let τ be a natural number greater than or equal to 1. Also, let a stochastic process $\{X_k\}$ be a discrete-time Markov chain whose state space and transition probability matrix, respectively, are $\{1, 2, \ldots, \tau + 1\}$ and

$$
\begin{bmatrix}
\rho_1 & 1-\rho_1 & 0 & \cdots & 0 \\
\rho_2 & 0 & 1-\rho_2 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\rho_\tau & 0 & 0 & \cdots & 1-\rho_\tau \\
1 & 0 & 0 & \cdots & 0
\end{bmatrix}.
$$

Proposition 1. Consider the above Markov chain $\{X_k\}$. If $\Pr\{X_0 = 1\} = 1$, then $\Pr\{X_k = 1\} \geq \rho_1^\tau$ for all $k \in \mathbb{Z}_{\geq 0}$.

Proof. We denote by $p_k^{(i)} := \Pr\{X_k = i\}$ the probability that the state is i at time k and prove $p_k^{(1)} \geq \rho_1^\tau$ by induction. Firstly, we consider the case in which the state transits within State 1. In this case, $p_t^{(1)} \geq p_0^{(1)}\rho_1^t \geq \rho_1^\tau$ $(t = 0, 1, \ldots, \tau)$ holds since $p_0^{(1)} = 1$.

Next, we assume all values less than $k+1 (> \tau)$ is true. We obtain the following simultaneous recurrence formulas:

$$
\begin{cases}
p_{k+1}^{(1)} = p_k^{(\tau+1)} + \sum_{i=1}^{\tau} \rho_i p_k^{(i)} = \sum_{i=1}^{\tau+1} \rho_i p_k^{(i)} \\
p_{k+1}^{(i)} = (1-\rho_{i-1})p_k^{(i-1)} \quad (i = 2, 3, \ldots, \tau+1)
\end{cases},
$$

where $\rho_{\tau+1} := 1$. Now,

$$
p_k^{(i)} = (1-\rho_{i-1})p_{k-1}^{(i-1)} = (1-\rho_{i-1})(1-\rho_{i-2})p_{k-2}^{(i-2)}
$$
$$
= \cdots = \Big\{ \prod_{j=1}^{i-1}(1-\rho_j) \Big\} p_{k-i+1}^{(1)},
$$

and thus, we obtain

$$
p_{k+1}^{(1)} = \sum_{i=1}^{\tau+1} \Big\{ \rho_i \prod_{j=1}^{i-1}(1-\rho_j) \Big\} p_{k-i+1}^{(1)}.
$$

Since the sum of the coefficient in the right-hand side of the above equation is 1, the assertion of the proposition holds at $k+1$ from the assumption of induction. $\square$

Appendix B. PROPERTY OF GAUSSIAN POLICY

Proposition 2. Let Assumptions 1 and 2 hold. For the state $x_k \in \mathbb{R}^n$ at time k of the nonlinear system given in (1), we select the input $u_k \in \mathbb{R}^m$ according to the m-dimensional normal distributed function with the mean $\mu_k \in \mathbb{R}^m$ and the covariance $\Sigma \in \mathbb{R}^{m \times m}$ given in (17). Then, the inequality condition

$$
\sigma \leq \frac{1}{\|h_j^\mathrm{T}B\|_2 \Phi^{-1}(q)} \left\{ v_j - h_j^\mathrm{T}(Ax_k + B\mu_k + \epsilon) \right\},
$$
$$
\forall j \text{ s.t. } h_j^\mathrm{T}B \neq 0, \ \forall \epsilon \in \mathcal{E}, \quad \text{(B.1)}
$$
$$
h_j^\mathrm{T}(Ax_k + \epsilon) \leq v_j, \forall j \text{ s.t. } h_j^\mathrm{T}B = 0, \ \forall \epsilon \in \mathcal{E} \quad \text{(B.2)}
$$

is the sufficient condition to satisfy $\mathsf{CC}'(x_{k+1}, q)$, where $\Phi(\cdot)$ is the cumulative distribution function of a zero mean unit variance Gaussian random variable, and $q \in (0.5, 1)$.

Proof. Let us denote the nominal value $\hat{x}_{k+1}$ at time $k+1$ of the nonlinear system by

$$
\hat{x}_{k+1} := Ax_k + Bu_k. \quad \text{(B.3)}
$$

With this value, we obtain

$$
\hat{x}_{k+1} \sim \mathcal{N}(Ax_k + B\mu_k, B^\mathrm{T}\Sigma B). \quad \text{(B.4)}
$$

Firstly, let us select an arbitrary j. In this case,

$$
\mathsf{CC}'(x_{k+1}, q) \Leftrightarrow \Pr\{h_j^\mathrm{T}x_{k+1} \leq v_j\} \geq q
$$
$$
\Leftarrow \Pr\{h_j^\mathrm{T}(\hat{x}_{k+1} + \epsilon) \leq v_j\} \geq q, \ \forall \epsilon \in \mathcal{E}.
$$

Next, let us select an arbitrary $\epsilon \in \mathcal{E}$. In this case, the above chance constraint becomes the following deterministic constraint (Boyd and Vandenberghe, 2004):

$$
\Pr\{h_j^\mathrm{T}(\hat{x}_{k+1} + \epsilon) \leq v_j\} \geq q
$$
$$
\Leftrightarrow v_j - h_j^\mathrm{T}(Ax_k + B\mu_k + \epsilon) \geq \Phi^{-1}(q)\left\| h_j^\mathrm{T}B\Sigma^{\frac{1}{2}} \right\|_2. \quad \text{(B.5)}
$$

Therefore, the inequality condition

$$
\Phi^{-1}(q)\|h_j^\mathrm{T}B\Sigma^{\frac{1}{2}}\|_2 \leq v_j - h_j^\mathrm{T}(Ax_k + B\mu_k + \epsilon),
$$
$$
\forall j = 1, 2, \ldots, n_c, \ \forall \epsilon \in \mathcal{E} \quad \text{(B.6)}
$$

is a sufficient condition for the state x_{k+1} to satisfy $\mathsf{CC}'(x_{k+1}, q)$ at time $k+1$. By dividing the above inequality condition depending on whether $h_j^\mathrm{T}B = 0$ or not, we obtain with $0 < \Phi^{-1}(q) < \infty$ that

$$
\|h_j^\mathrm{T}B\Sigma^{\frac{1}{2}}\|_2 \leq \frac{1}{\Phi^{-1}(q)}\left\{ v_j - h_j^\mathrm{T}(Ax_k + B\mu_k + \epsilon) \right\}
$$
$$
\forall j \text{ s.t. } h_j^\mathrm{T}B \neq 0, \ \forall \epsilon \in \mathcal{E}, \quad \text{(B.7)}
$$
$$
h_j^\mathrm{T}(Ax_k + \epsilon) \leq v_j, \forall j \text{ s.t. } h_j^\mathrm{T}B = 0, \ \forall \epsilon \in \mathcal{E}. \quad \text{(B.8)}
$$

In addition, with the variance-covariance matrix given in (17), we obtain

$$
\left\| h_j^\mathrm{T}B\Sigma^{\frac{1}{2}} \right\|_2 = \sigma\|h_j^\mathrm{T}B\|_2. \quad \text{(B.9)}
$$

Therefore, by substituting this into (B.7) with respect to j which satisfies $h_j^\mathrm{T}B \neq 0$ and dividing its both side by $\|h_j^\mathrm{T}B\|_2$, we obtain the condition in (B.1). $\square$

Corollary 3. Let assumptions used in Proposition 2 hold. In addition, Assumption 5 holds. Then,

$$
\sigma \leq \frac{1}{\|h_j^\mathrm{T}B\|_2 \Phi^{-1}(q)} \left\{ v_j - h_j^\mathrm{T}(Ax_k + B\mu_k + \epsilon) \right\}
$$
$$
\forall j = 1, 2, \ldots, n_c, \ \forall \epsilon \in \mathcal{E} \quad \text{(B.10)}
$$

is a sufficient condition to satisfy $\mathsf{CC}'(x_{k+1}, q)$.

Proof. We can prove this corollary from Proposition 2 and the assumption $h_j^\mathrm{T}B \neq 0, \forall j = 1, 2, \ldots, n_c$. $\square$