SAFE EXPLORATION METHOD FOR REINFORCEMENT LEARNING UNDER EXISTENCE OF DISTURBANCE

Yoshihiro Okawa[1*], **Yusuke Kato**[2*], **Tomotake Sasaki**[1], **Hitoshi Yanami**[1], **Toru Namerikawa**[2]
[1] Fujitsu Laboratories Ltd., [2] Keio University
{okawa.y,tomotake.sasaki,yanami}@fujitsu.com, namerikawa@keio.jp

ABSTRACT

In this study, we deal with a safe exploration problem in reinforcement learning under existence of disturbance. We define the safety during learning as satisfaction of state constraints defined explicitly in the problems and propose an automatic exploration process adjustment method that uses prior knowledge of a controlled object and disturbance. The proposed method assures the satisfaction of the explicit state constraints with a pre-specified probability even if the controlled object is exposed to a stochastic disturbance following a normal distribution. We evaluate the validity and effectiveness of our method through a numerical simulation.

1 INTRODUCTION

Guaranteeing safety and performance during learning is one of the critical issues to implement reinforcement learning (RL) in real environments. To address this issue, RL algorithms and related methods dealing with safety have been studied in recent years (García & Fernández, 2015). For example, Biyik et al. (2019) proposed a safe exploration algorithm used in RL. They guaranteed to prevent states from being unrecoverable by leveraging the Lipschitz-continuity of its unknown transition model. In addition, Ge et al. (2019) proposed a modified Q-learning method for a constrained Markov decision process (MDP) to seek for the optimal solution ensuring that the safety premise is satisfied. However, few studies evaluated their safety quantitatively from a viewpoint of satisfying state constraints at each time that are defined explicitly in the problems. Recently, Okawa et al. (2020) proposed an automatic exploration process adjustment method that is applicable to existing RL algorithms. They quantitatively evaluated the above-mentioned safety in accordance with probabilities of satisfying the explicit state constraints. However, they did not consider the existence of external disturbance, which is an important factor when we consider safety. In particular, such disturbance sometimes makes the states violate their constraints even if the inputs used in exploration are designed to satisfy those constraints. In this study, we tackle the safe exploration problem in RL under the existence of the disturbance. We extend the previous work (Okawa et al., 2020) and propose an automatic exploration process adjustment method for RL that uses partially known information of both the controlled object and disturbance to guarantee the safety during learning. The proposed method assures the satisfaction of explicit state constraints with a pre-specified probability even if the controlled object is exposed to a disturbance following a normal distribution. We show the effectiveness of the proposed method with a numerical simulation.

2 PROBLEM FORMULATION

We consider an input-affine discrete-time nonlinear dynamic system written in the following form:

$$\boldsymbol{x}_{k+1} = \boldsymbol{f}(\boldsymbol{x}_k) + \boldsymbol{G}(\boldsymbol{x}_k)\boldsymbol{u}_k + \boldsymbol{w}_k, \tag{1}$$

where $\boldsymbol{x}_k \in \mathbb{R}^n$, $\boldsymbol{u}_k \in \mathbb{R}^m$, and $\boldsymbol{w}_k \in \mathbb{R}^n$ stand for the state, input and disturbance at time k, respectively. In addition, $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n$ and $\boldsymbol{G} : \mathbb{R}^n \to \mathbb{R}^{n\times m}$ are unknown nonlinear functions. We suppose the state $\boldsymbol{x}_k$ is directly observable. An immediate cost $c_{k+1} \geq 0$ is given depending on the state, input and disturbance at each time k:

$$c_{k+1} = c(\boldsymbol{x}_k,\ \boldsymbol{u}_k,\ \boldsymbol{w}_k), \tag{2}$$

[*] Y. Okawa and Y. Kato contributed equally to this work.

where the immediate cost function $c : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n \to [0, \infty)$ is unknown while c_{k+1} is supposed to be directly observable. We consider the situation where the constraints that the state should satisfy from the viewpoint of safety are explicitly given by the following linear inequalities:

$$\boldsymbol{H}\boldsymbol{x} \preceq \boldsymbol{d}, \tag{3}$$

where $\boldsymbol{d} = [d_1, \ldots, d_{n_c}]^\top \in \mathbb{R}^{n_c}$, $\boldsymbol{H} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_{n_c}]^\top \in \mathbb{R}^{n_c \times n}$, n_c is the number of constraints and $\preceq$ means that the inequality $\leq$ holds for all elements. In addition, we define $\mathcal{X}_s \subset \mathbb{R}^n$ as the set of safe states, that is,

$$\mathcal{X}_s := \{\boldsymbol{x} \in \mathbb{R}^n | \boldsymbol{H}\boldsymbol{x} \preceq \boldsymbol{d}\}. \tag{4}$$

Initial state $\boldsymbol{x}_0$ is assumed to satisfy $\boldsymbol{x}_0 \in \mathcal{X}_s$ for simplicity.

The goal of reinforcement learning is to acquire a policy (control law) that minimizes or maximizes an evaluation function with respect to the immediate cost or reward, using them as cues in its trial-and-error process. In this study, we consider the standard discounted cumulative cost as the evaluation function to be minimized:

$$J = \sum_{k=0}^{T-1} \gamma^k c_{k+1}. \tag{5}$$

Here, γ is a discount factor ($0 < \gamma \leq 1$) and T is the terminal time.

Besides (5) for the cost evaluation, we define the safety in this problem as satisfaction of the state constraints and evaluate its guarantee quantitatively. In detail, we consider the following chance constraint with respect to the satisfaction of the explicit state constraints (3) at each time:

$$\Pr\{\boldsymbol{H}\boldsymbol{x}_k \preceq \boldsymbol{d}\} \geq \eta, \tag{6}$$

where $\Pr\{\boldsymbol{H}\boldsymbol{x}_k \preceq \boldsymbol{d}\} (= \Pr\{\boldsymbol{x}_k \in \mathcal{X}_s\})$ denotes the probability that $\boldsymbol{x}_k$ satisfies the constraints (3).

The objective of the proposed method is to ensure that the chance constraint (6) is satisfied at every time $k = 1, 2, \ldots, T$ for a pre-specified η, which is $0.5 < \eta < 1$ in this study.

Figure 1 shows the overall picture of the reinforcement learning problem in this study. The controller generates an input $\boldsymbol{u}_k$ according to the proposed method and apply it to the controlled object, which is a discrete-time nonlinear dynamic system exposed to a disturbance $\boldsymbol{w}_k$. The proposed method includes a base policy, and it is updated based on the states $\boldsymbol{x}_{k+1}$ and immediate cost c_{k+1} observed from the controlled object. In addition to update the base policy to minimize the evaluation function, the chance constraint should be satisfied at every time k.

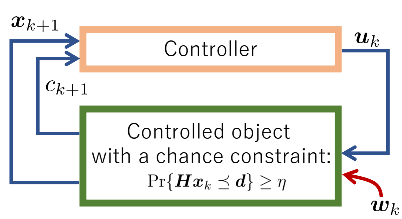


Figure 1: Overview of controller and controlled object under existence of disturbance

As the base policy, we consider the Gaussian probability density function of the following form:

$$\Pi(\boldsymbol{u}|\boldsymbol{x};\boldsymbol{\omega};\boldsymbol{\Sigma}) = \frac{1}{(\sqrt{2\pi})^m \sqrt{|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}\left(\boldsymbol{u} - \boldsymbol{\mu}(\boldsymbol{x};\boldsymbol{\omega})\right)^\top \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{u} - \boldsymbol{\mu}(\boldsymbol{x};\boldsymbol{\omega})\right)\right), \tag{7}$$

where $\boldsymbol{\mu}(\boldsymbol{x};\boldsymbol{\omega}) \in \mathbb{R}^m$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{m \times m}$ are the mean and the variance-covariance matrix, respectively. We notate generating inputs probabilistically according to this Gaussian probability density function as $\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{x};\boldsymbol{\omega}), \boldsymbol{\Sigma})$. We denote the mean of input at time k as $\boldsymbol{\mu}_k := \boldsymbol{\mu}(\boldsymbol{x}_k;\boldsymbol{\omega}_k)$, which is determined by the state $\boldsymbol{x}_k$ and the policy parameter $\boldsymbol{\omega}_k \in \mathbb{R}^{N_{\boldsymbol{\omega}}}$.

We make the following six assumptions. The proposed method uses the partial information of the controlled object and the disturbance.

Assumption 1. *Matrices $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{B} \in \mathbb{R}^{n\times m}$ in the following linear approximation model of the nonlinear system (1) are known:*

$$\boldsymbol{x}_{k+1} \simeq \boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{u}_k + \boldsymbol{w}_k. \tag{8}$$

The next assumption is about the disturbance.

Assumption 2. *Disturbance $\boldsymbol{w}_k$ stochastically occur according to an n-dimensional normal distribution represented by the following Gaussian probability density function:*

$$\Pi(\boldsymbol{w}) = \frac{1}{(\sqrt{2\pi})^n\sqrt{|\boldsymbol{\Sigma}_w|}}\exp\left(-\frac{1}{2}\left(\boldsymbol{w}-\boldsymbol{\mu}_w\right)^\top \boldsymbol{\Sigma}_w^{-1}\left(\boldsymbol{w}-\boldsymbol{\mu}_w\right)\right), \tag{9}$$

where $\boldsymbol{\mu}_w \in \mathbb{R}^n$ and $\boldsymbol{\Sigma}_w \in \mathbb{R}^{n\times n}$ are the mean and the variance-covariance matrix, respectively. We notate this as $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)$. Also, the mean $\boldsymbol{\mu}_w$ and variance-covariance matrix $\boldsymbol{\Sigma}_w$ are known, and disturbance $\boldsymbol{w}_k$ and input $\boldsymbol{u}_k$ are uncorrelated.

The following assumption about the linear approximation model and the constraints is also made.

Assumption 3. *The following condition holds for $\boldsymbol{B}$ and $\boldsymbol{H} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_{n_c}]^\top$:*

$$\boldsymbol{h}_j^\top \boldsymbol{B} \neq \boldsymbol{0}, \quad \forall j = 1, 2, \ldots, n_c. \tag{10}$$

In order to achieve the aforementioned objective, we use inputs that do not contain exploring aspect depending on the state at each time. The following two assumptions are regarding the input without exploring aspect, which we refer to as conservative inputs.

Assumption 4. *Suppose $\boldsymbol{x}_k = \boldsymbol{x} \notin \mathcal{X}_s$ occurs at time $k \geq 1$. Input sequence $\boldsymbol{u}_k^{back}$, $\boldsymbol{u}_{k+1}^{back}$, $\ldots$, $\boldsymbol{u}_{k+j-1}^{back}$ or how to generate them is known, such that the probability of the state moving back to $\mathcal{X}_s$ within τ ($\tau < T$) steps is greater than or equal to ξ ($\eta^{\frac{1}{T}} < \xi < 1$), regardless of time k and state $\boldsymbol{x}$. In other words, using known input sequence $\boldsymbol{u}_k^{back}$, $\boldsymbol{u}_{k+1}^{back}$, $\ldots$, $\boldsymbol{u}_{k+j-1}^{back}$ for some $j \leq \tau$, $\Pr\{\boldsymbol{x}_{k+j} \in \mathcal{X}_s\} \geq \xi$ holds.*

Assumption 5. *Suppose $\boldsymbol{x}_k = \boldsymbol{x} \in \mathcal{X}_s$ occurs at time $k \geq 0$. Input $\tilde{\boldsymbol{u}}_k$ or how to generate it is known, such that $\Pr\{\boldsymbol{H}\boldsymbol{x}_{k+1} \preceq \boldsymbol{d}\} \geq \left(\frac{\eta}{\xi^k}\right)^{\frac{1}{\tau}}$ holds.*

We define the difference (i.e., approximation error) $\boldsymbol{e}(\boldsymbol{x}, \boldsymbol{u}) = [e_1(\boldsymbol{x}, \boldsymbol{u}),\ \ldots,\ e_n(\boldsymbol{x}, \boldsymbol{u})]^\top \in \mathbb{R}^n$ between the nonlinear system (1) and the linear approximation model (8) as below:

$$\boldsymbol{e}(\boldsymbol{x}, \boldsymbol{u}) = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{G}(\boldsymbol{x})\boldsymbol{u} - (\boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u}). \tag{11}$$

We make the following assumption.

Assumption 6. *Regarding the approximation error $\boldsymbol{e}(\boldsymbol{x}, \boldsymbol{u})$ expressed as (11), $\bar{\delta}_j < \infty$, $\bar{\hat{\delta}}_j < \infty$, $j = 1, 2, \ldots, n_c$ that satisfy the following inequalities are known:*

$$\bar{\delta}_j \geq \sup_{\boldsymbol{x}\in\mathbb{R}^n,\ \boldsymbol{u}\in\mathbb{R}^m} |\boldsymbol{h}_j^\top \boldsymbol{e}(\boldsymbol{x}, \boldsymbol{u})|, \quad j = 1, 2, \ldots, n_c, \tag{12}$$

$$\bar{\hat{\delta}}_j \geq \sup_{\boldsymbol{x}\in\mathbb{R}^n,\ \boldsymbol{u}\in\mathbb{R}^m} |\boldsymbol{h}_j^\top \left(\boldsymbol{A}^{\tau-1} + \boldsymbol{A}^{\tau-2} + \cdots + \boldsymbol{I}\right)\boldsymbol{e}(\boldsymbol{x}, \boldsymbol{u})|, \quad j = 1, 2, \ldots, n_c. \tag{13}$$

3 Exploration process adjustment with conservative inputs

To guarantee the safety with respect to the satisfaction of the chance constraint (6), the following is the automatic exploration adjustment method we propose:

$$\begin{cases} \text{(i) } \boldsymbol{u}_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\ \quad \text{if } \boldsymbol{x}_k \in \mathcal{X}_s \wedge \left(\left\|\boldsymbol{h}_j^\top \boldsymbol{\Sigma}_w^{\frac{1}{2}}\right\|_2 \leq \frac{1}{\Phi^{-1}(\eta_k')}(d_j - \boldsymbol{h}_j^\top \hat{\boldsymbol{x}}_{k+1} - \delta_j), \forall \delta_j \in \{\pm\bar{\delta}_j\}, \forall j = 1, \ldots, n_c\right), \\ \text{(ii) } \boldsymbol{u}_k = \tilde{\boldsymbol{u}}_k \\ \quad \text{if } \boldsymbol{x}_k \in \mathcal{X}_s \wedge \left(\left\|\boldsymbol{h}_j^\top \boldsymbol{\Sigma}_w^{\frac{1}{2}}\right\|_2 > \frac{1}{\Phi^{-1}(\eta_k')}(d_j - \boldsymbol{h}_j^\top \hat{\boldsymbol{x}}_{k+1} - \delta_j),\ \text{for some } \delta_j \in \{\pm\bar{\delta}_j\}\right), \\ \text{(iii) } \boldsymbol{u}_k = \boldsymbol{u}_k^{back} \text{ if } \boldsymbol{x}_k \notin \mathcal{X}_s, \end{cases} \tag{14}$$

where

$$\hat{\boldsymbol{x}}_{k+1} := \boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{\mu}_k + \boldsymbol{\mu}_w, \quad \eta'_k := 1 - \frac{1-\left(\frac{\eta}{\xi^k}\right)^{\frac{1}{\tau}}}{n_c}. \tag{15}$$

Here, $\tilde{\boldsymbol{u}}_k$ and $\boldsymbol{u}_k^{back}$ are the conservative inputs explained in the previous section. That is, this method switches the exploratory inputs and the conservative ones in accordance with the current and predicted state information by using prior knowledge of both the controlled object and disturbance, while the previous work (Okawa et al., 2020) only used that of the controlled object. In addition, with those prior knowledge, this method adjusts the degree of its exploration by calculating the variance-covariance matrix $\boldsymbol{\Sigma}_k$ online, which is a feasible solution for the following inequality:

$$\left\| \boldsymbol{h}_j^\top \boldsymbol{B}' \begin{bmatrix} \boldsymbol{\Sigma}_k & \\ & \boldsymbol{\Sigma}_w \end{bmatrix}^{\frac{1}{2}} \right\|_2 \leq \frac{1}{\Phi^{-1}(\eta'_k)}(d_j - \boldsymbol{h}_j^\top \hat{\boldsymbol{x}}_{k+1} - \delta_j), \forall \delta_j \in \{\pm\bar{\delta}_j\}, \forall j = 1, \ldots, n_c. \tag{16}$$

A more detailed explanation of this method is described in Appendix A. Under Assumptions 1–6, this method makes the chance constraint (6) satisfied at every time k even with the existence of disturbance. The theoretical guarantee is given in Appendix A as Theorem 1. We also provide another theoretical result (Theorem 2) regarding the construction of conservative inputs.

We have combined our proposed method with the Deep Deterministic Policy Gradient (DDPG) algorithm (Lillicrap et al., 2015) and tested its effectiveness on the simulated inverted-pendulum provided in OpenAI Gym (Brockman et al., 2016)[1]. Figure 2 shows the relative frequencies of constraint satisfaction regarding the angular velocity of the pendulum at each time k. We obtained these results with 100 episodes $\times$ 10 runs of the simulation (each episode consists of 100 time steps, during which the pendulum is tried to be swung up and held at the top). The red crosses denote the results with the proposed method and the green dashed line denotes the pre-specified probability η. For the reference, the results with the previous work (Okawa et al., 2020) not taking the disturbance into account are shown by blue triangles. We can see that the proposed method works as expected.

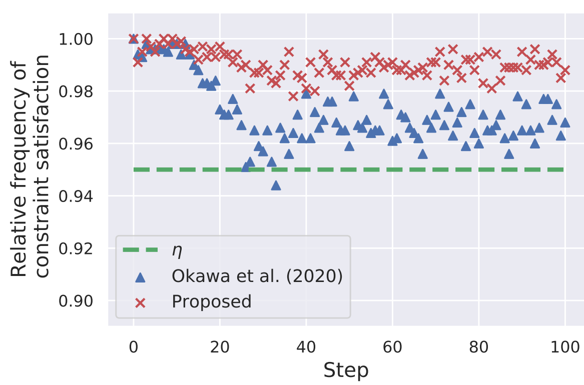


Figure 2: Relative frequencies of constraint satisfaction at each time step

Further details of this simulation are given in Appendix B.

4 Conclusion

In this study, we proposed an automatic exploration process adjustment method for RL to guarantee the safety during learning under the existence of the disturbances. The proposed method uses partially known information of both the controlled object and disturbances. As a result, this method assures the satisfaction of explicit state constraints with a pre-specified probability at every time even if the controlled object is exposed to the disturbance following a normal distribution. We showed the effectiveness of the proposed method with a numerical simulation.

[1] We have made a small modification of the code to add the disturbance.

References

E. Biyik, J. Margoliash, S. R. Alimo, and D. Sadigh. Efficient and safe exploration in deterministic Markov decision processes with unknown transition models. In *Proceedings of the American Control Conference (ACC)*, pp. 1792–1799, 2019.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. OpenAI Gym. *arXiv preprint arXiv:1606.01540*, 2016.

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

Y. Ge, F. Zhu, X. Ling, and Q. Liu. Safe Q-learning method based on constrained Markov decision processes. *IEEE Access*, 7:165007–165017, 2019.

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.

Y. Okawa, T. Sasaki, and H. Iwane. Automatic exploration process adjustment for safe reinforcement learning with joint chance constraint satisfaction. In *Proceedings of the 21st IFAC World Congress, IFAC-PapersOnLine*, 53(2):1588–1595, 2020. URL `https://www.sciencedirect.com/science/article/pii/S2405896320328524`.

Appendix

A Details of proposed method and theoretical guarantee for chance constraint satisfaction

The proposed method (14) generates inputs differently in accordance with the following three cases (Fig. 3): (i) the state constraints are satisfied and the input contains exploring aspect, (ii) the state constraints are satisfied but the input do not contain exploring aspect, and (iii) the state constraints are not satisfied. We consider the case (i) in Subsection A.1 and the case (iii) in Subsection A.2, respectively. Then, in Subsection A.3, we prove that the proposed method makes the chance constraint (6) satisfied at every time k under Assumptions 1–6 (Theorem 1). We also provide Theorem 2 regarding the construction of conservative inputs used in the cases (ii) and (iii) in Subsection A.4.

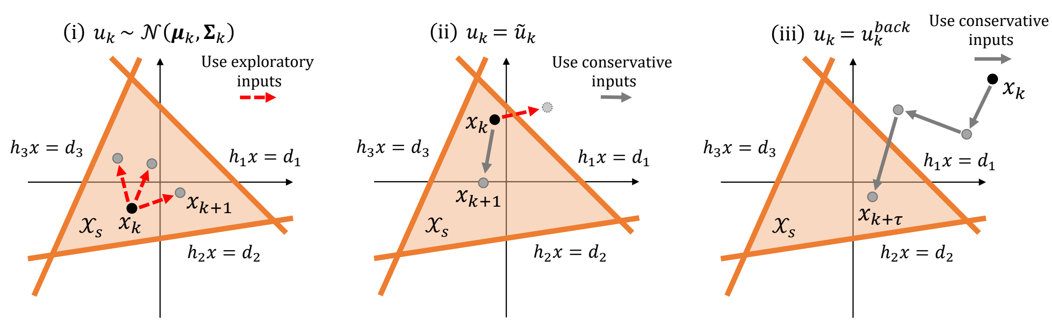


Figure 3: Illustration of the proposed method for a case of $n = 2$ and $n_c = 3$.

A.1 Adjusting inputs when using a Gaussian policy

First, we consider the case when we generate inputs containing exploring aspect according to the Gaussian probability density function (7). The following lemma holds.

Lemma 1. *Let $q \in (0.5,\ 1)$. Suppose Assumptions 1, 2, and 6 hold. Choose input $\boldsymbol{u}_k$ according to an m-dimensional Gaussian distribution with mean $\boldsymbol{\mu}_k$ and variance-covariance matrix $\boldsymbol{\Sigma}_k$ when the state of the nonlinear system (1) at time k is $\boldsymbol{x}_k$. Then, the following inequality is a sufficient condition for $\Pr\{\boldsymbol{h}_j^\top \boldsymbol{x}_{k+1} \leq d_j\} \geq q,\ \forall j = 1, \ldots, n_c$:*

$$\left\| \boldsymbol{h}_j^\top \boldsymbol{B}' \begin{bmatrix} \boldsymbol{\Sigma}_k & \\ & \boldsymbol{\Sigma}_w \end{bmatrix}^{\frac{1}{2}} \right\|_2 \leq \frac{1}{\Phi^{-1}(q)} \left\{ d_j - \boldsymbol{h}_j^\top (\boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{\mu}_k + \boldsymbol{\mu}_w) + \delta_j \right\}, \quad \forall j = 1,\ 2,\ \ldots,\ n_c,\ \ \forall \delta_j \in \{\bar{\delta}_j,\ -\bar{\delta}_j\}, \tag{17}$$

where $\boldsymbol{B}' = [\boldsymbol{B}, \boldsymbol{I}]$ and $\Phi(\cdot)$ is the normal cumulative distribution function.

Proof. By using the definition of $\boldsymbol{e}$, that is (11), we can rewrite the state equation (1) as follows:

$$\boldsymbol{x}_{k+1} = \boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{u}_k + \boldsymbol{e}(\boldsymbol{x}_k, \boldsymbol{u}_k) + \boldsymbol{w}_k. \tag{18}$$

If one j is arbitrarily selected and fixed, the following relation holds for the state $\boldsymbol{x}_{k+1}$:

$$\begin{aligned}
&\Pr\{\boldsymbol{h}_j^\top \boldsymbol{x}_{k+1} \leq d_j\} \geq q \\
&\Leftrightarrow \Pr\{\boldsymbol{h}_j^\top (\boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{u}_k + \boldsymbol{e}(\boldsymbol{x}_k, \boldsymbol{u}_k) + \boldsymbol{w}_k) \leq d_j\} \geq q \\
&\Leftarrow \Pr\{\boldsymbol{h}_j^\top (\boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{u}_k + \boldsymbol{w}_k) + \delta_j \leq d_j\} \geq q,\ \forall \delta_j \in \{\bar{\delta}_j,\ -\bar{\delta}_j\} \quad \text{(due to Assumption 6)} \\
&\Leftrightarrow \Pr\{\boldsymbol{h}_j^\top (\boldsymbol{A}\boldsymbol{x}_k + [\boldsymbol{B}, \boldsymbol{I}] \begin{bmatrix} \boldsymbol{u}_k \\ \boldsymbol{w}_k \end{bmatrix}) + \delta_j \leq d_j\} \geq q,\ \forall \delta_j \in \{\bar{\delta}_j,\ -\bar{\delta}_j\} \\
&\Leftrightarrow \Pr\{\boldsymbol{h}_j^\top (\boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}' \begin{bmatrix} \boldsymbol{u}_k \\ \boldsymbol{w}_k \end{bmatrix}) + \delta_j \leq d_j\} \geq q,\ \forall \delta_j \in \{\bar{\delta}_j,\ -\bar{\delta}_j\}.
\end{aligned} \tag{19}$$

Input $\boldsymbol{u}_k$ and disturbance $\boldsymbol{w}_k$ are uncorrelated (Assumption 2), so if one $\delta_j \in \{\bar{\delta}_j,\ -\bar{\delta}_j\}$ is arbitrarily selected and fixed, the following relation holds (Boyd & Vandenberghe, 2004):

$$\begin{aligned}
&\Pr\{\boldsymbol{h}_j^\top (\boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}' \begin{bmatrix} \boldsymbol{u}_k \\ \boldsymbol{w}_k \end{bmatrix}) + \delta_j \leq d_j\} \geq q \\
&\Leftrightarrow d_j - \boldsymbol{h}_j^\top \left(\boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}' \begin{bmatrix} \boldsymbol{\mu}_k \\ \boldsymbol{\mu}_w \end{bmatrix} \right) - \delta_j \geq \Phi^{-1}(q) \left\| \boldsymbol{h}_j^\top \boldsymbol{B}' \begin{bmatrix} \boldsymbol{\Sigma}_k & \\ & \boldsymbol{\Sigma}_w \end{bmatrix}^{\frac{1}{2}} \right\|_2.
\end{aligned} \tag{20}$$

Hence,

$$\begin{aligned}
&\Phi^{-1}(q) \left\| \boldsymbol{h}_j^\top \boldsymbol{B}' \begin{bmatrix} \boldsymbol{\Sigma}_k & \\ & \boldsymbol{\Sigma}_w \end{bmatrix}^{\frac{1}{2}} \right\|_2 \leq d_j - \boldsymbol{h}_j^\top (\boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{\mu}_k + \boldsymbol{\mu}_w) - \delta_j \\
&\qquad \forall j = 1,\ 2,\ \ldots,\ n_c,\ \ \forall \delta_j \in \{\bar{\delta}_j,\ -\bar{\delta}_j\} \\
&\Rightarrow \Pr\{\boldsymbol{h}_j^\top \boldsymbol{x}_{k+1} \leq d_j\} \geq q,\ \forall j = 1,\ 2,\ \ldots,\ n_c.
\end{aligned} \tag{21}$$

Note that $\Phi^{-1}(q) > 0$ for $q \in (0.5,\ 1)$. Thus, the inequality on the left side of (21) can be rewritten as follows:

$$\left\| \boldsymbol{h}_j^\top \boldsymbol{B}' \begin{bmatrix} \boldsymbol{\Sigma}_k & \\ & \boldsymbol{\Sigma}_w \end{bmatrix}^{\frac{1}{2}} \right\|_2 \leq \frac{1}{\Phi^{-1}(q)} \left\{ d_j - \boldsymbol{h}_j^\top (\boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{\mu}_k + \boldsymbol{\mu}_w) - \delta_j \right\}, \quad \forall j = 1,\ 2,\ \ldots,\ n_c,\ \ \forall \delta_j \in \{\bar{\delta}_j,\ -\bar{\delta}_j\} \tag{22}$$

□

A.2 Adjusting inputs when state constraints are violated

Next, we consider the case when the state constraints are not satisfied. In this case, we use the input sequence considered in Assumption 4. Regarding this situation, the following lemma holds.

Lemma 2. *Suppose we use input sequence $\boldsymbol{u}_k^{back},\ \boldsymbol{u}_{k+1}^{back},\ \ldots,\ \boldsymbol{u}_{k+j-1}^{back}$ ($j < \tau$) considered in Assumption 4 when $\boldsymbol{x}_{k-1} \in \mathcal{X}_s$ and $\boldsymbol{x}_k = \boldsymbol{x} \notin \mathcal{X}_s$ occur. Also suppose $\boldsymbol{x}_k \in \mathcal{X}_s \Rightarrow \Pr\{\boldsymbol{x}_{k+1} \in \mathcal{X}_s\} \geq p$ holds with a $p \in (0, 1)$. Then $\Pr\{\boldsymbol{x}_k \in \mathcal{X}_s\} \geq \xi^k p^\tau$ holds for all $k = 1, 2, \ldots, T$ if $\boldsymbol{x}_0 \in \mathcal{X}_s$.*

Proof. Consider a discrete-time Markov chain $\{X_k\}$, where $\{1, 2, \ldots, \tau, \tau+1, \tau+2\}$ is the state space and the transition probability matrix is as follows:

$$\begin{bmatrix} \rho_1 & 1-\rho_1 & 0 & \cdots & 0 & 0 \\ \rho_2 & 0 & 1-\rho_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_\tau & 0 & 0 & \cdots & 1-\rho_\tau & 0 \\ \rho_{\tau+1} & 0 & 0 & \cdots & 0 & 1-\rho_{\tau+1} \\ \rho_{\tau+2} & 0 & 0 & \cdots & 0 & 1-\rho_{\tau+2} \end{bmatrix}. \tag{23}$$

The state transition diagram is shown in Figure 4.

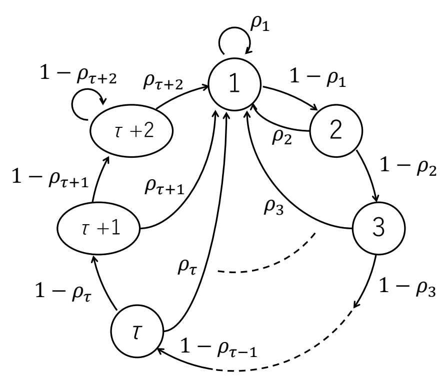


Figure 4: State transition diagram

We use this Markov chain to prove the lemma by relating this to our main problem as follows: "$\boldsymbol{x}_k \in \mathcal{X}_s$" is state 1, "$\boldsymbol{x}_{k-i} \in \mathcal{X}_s$ and $\boldsymbol{x}_{k-i+1}, \ldots, \boldsymbol{x}_k \notin \mathcal{X}_s$" is state $i+1$ ($i = 1, 2, \ldots, \tau$), and "$\boldsymbol{x}_{k-\tau}, \boldsymbol{x}_{k-\tau+1}, \ldots, \boldsymbol{x}_k \notin \mathcal{X}_s$" is state $\tau+2$. We define the probability of the state of Markov chain being i at time k as $p_k^{(i)}$, that is,

$$p_k^{(i)} := \Pr\{X_k = i\}. \tag{24}$$

We prove by induction that the inequality

$$p_k^{(1)} \geq \xi^k \rho_1^\tau \tag{25}$$

holds for all $k = 1, 2, \ldots, T$ when $p_0^{(1)} = \Pr\{X_0 = 1\} = \Pr\{\boldsymbol{x}_0 \in \mathcal{X}_s\} = 1$.

First, consider $k = 1, 2, \ldots, \tau$. We have the following relation:

$$\begin{aligned} p_k^{(1)} &\geq p_0^{(1)} \rho_1^k \\ &= \rho_1^k \qquad (\text{because } p_0^{(1)} = 1) \\ &\geq \rho_1^\tau \qquad (\text{because } k \leq \tau) \\ &\geq \xi^k \rho_1^\tau. \end{aligned} \tag{26}$$

Next, consider $k = \tau + 1$. From the following relations,

$$p_{\tau+1}^{(1)} = \sum_{i=1}^{\tau+1} \rho_i p_\tau^{(i)}, \tag{27}$$

$$p_\tau^{(i)} = \left(\prod_{j=1}^{i-1} (1-\rho_j) \right) p_{\tau-i+1}^{(1)} \; (i = 2, \ldots, \tau+1), \tag{28}$$

we have the following:

$$
\begin{aligned}
p_{\tau+1}^{(1)} &= \sum_{i=1}^{\tau+1} \rho_i p_\tau^{(i)} \\
&= \rho_1 p_\tau^{(1)} + \sum_{i=2}^{\tau+1} \rho_i \left(\prod_{j=1}^{i-1} (1-\rho_j) \right) p_{\tau-i+1}^{(1)} \\
&\geq \rho_1 \xi^\tau \rho_1^\tau + \sum_{i=2}^{\tau+1} \rho_i \left(\prod_{j=1}^{i-1} (1-\rho_j) \right) \xi^\tau \rho_1^\tau \\
&= \left\{ \rho_1 + \sum_{i=2}^{\tau+1} \rho_i \prod_{j=1}^{i-1} (1-\rho_j) \right\} \xi^\tau \rho_1^\tau \\
&= \left\{ \rho_1 + (1-\rho_1) \left(\rho_2 + \sum_{i=3}^{\tau+1} \rho_i \prod_{j=2}^{i-1} (1-\rho_j) \right) \right\} \xi^\tau \rho_1^\tau.
\end{aligned} \tag{29}
$$

From Assumption 4, the probability of moving inside $\mathcal{X}_s$ within τ steps is greater than or equal to ξ if we use $\boldsymbol{u}_k^{back},\ \boldsymbol{u}_{k+1}^{back},\ \dots$ when $\boldsymbol{x}_{k-1} \in \mathcal{X}_s$ and $\boldsymbol{x}_k \notin \mathcal{X}_s$ occur. Thus the following inequality holds:

$$
\rho_2 + \sum_{i=3}^{\tau+1} \rho_i \prod_{j=2}^{i-1} (1-\rho_j) \geq \xi. \tag{30}
$$

From (29) and (30), we have

$$
\begin{aligned}
p_{\tau+1}^{(1)} &\geq \{\rho_1 + (1-\rho_1)\xi\} \xi^\tau \rho_1^\tau \\
&= \{\xi + \rho_1(1-\xi)\} \xi^\tau \rho_1^\tau \\
&\geq \xi^{\tau+1} \rho_1^\tau.
\end{aligned} \tag{31}
$$

Therefore, $p_k^{(1)} \geq \xi^k \rho_1^\tau$ holds also at $k = \tau+1$.

Suppose that $p_k^{(1)} \geq \xi^k \rho_1^\tau$ holds for $k \geq \tau+1$. The following recurrence formulas hold:

$$
\begin{cases}
p_{k+1}^{(1)} = \sum_{i=1}^{\tau+2} \rho_i p_k^{(i)}, \\
p_{k+1}^{(i)} = (1-\rho_{i-1}) p_k^{(i-1)} \quad (i = 2, 3, \dots, \tau+1), \\
p_{k+1}^{(\tau+2)} = (1-\rho_{\tau+1}) p_k^{(\tau+1)} + (1-\rho_{\tau+2}) p_k^{(\tau+2)}.
\end{cases} \tag{32}
$$

From

$$
p_k^{(i)} = \left(\prod_{j=1}^{i-1} (1-\rho_j) \right) p_{k-i+1}^{(1)}, \tag{33}
$$

we obtain the following relation:

$$\begin{aligned}
p_{k+1}^{(1)} &= \sum_{i=1}^{\tau+2} \rho_i p_k^{(i)} \\
&\geq \sum_{i=1}^{\tau+1} \rho_i p_k^{(i)} \\
&= \rho_1 p_k^{(1)} + \sum_{i=2}^{\tau+1} \rho_i \left(\prod_{j=1}^{i-1} (1-\rho_j) \right) p_{k-i+1}^{(1)} \\
&\geq \rho_1 \xi^k \rho_1^{\tau+1} + \sum_{i=2}^{\tau+1} \rho_i \left(\prod_{j=1}^{i-1} (1-\rho_j) \right) \xi^k \rho_1^{\tau} \\
&= \left\{ \rho_1 + \sum_{i=2}^{\tau+1} \rho_i \prod_{j=1}^{i-1} (1-\rho_j) \right\} \xi^k \rho_1^{\tau} \\
&\geq \{\rho_1 + (1-\rho_1)\xi\} \xi^k \rho_1^{\tau} \\
&= \{\xi + \rho_1(1-\xi)\} \xi^k \rho_1^{\tau} \\
&\geq \xi^{k+1} \rho_1^{\tau}.
\end{aligned} \tag{34}$$

Hence, $p_k^{(1)} \geq \xi^k \rho_1^\tau$ holds also at $k+1$.

Therefore, $p_k^{(1)} \geq \xi^k \rho_1^\tau$ hold for all $k = 1, 2, \ldots, T$. Note that $\Pr\{\boldsymbol{x}_1 \in \mathcal{X}_s\} = \rho_1$ because $\Pr\{\boldsymbol{x}_0 \in \mathcal{X}_s\} = 1$, and thus $\rho_1 \geq p$. Now the lemma is proved. □

A.3 Theoretical guarantee for chance constraint satisfaction

Under Assumptions 1–6, the proposed method (14) makes the chance constraint (6) satisfied at every time k even with the existence of disturbance. We give a theoretical guarantee below.

Theorem 1. *Let $\eta \in (0.5, 1)$. Suppose Assumptions 1 through 6 hold. Then, by determining input $\boldsymbol{u}_k$ according to the proposed method (14), chance constraints (6) are satisfied at all time $k = 1, 2, \ldots, T$.*

Proof. First, consider the case of (i) in (14). Remember that ξ is a scalar considered in Assumption 4 such that $\eta^{\frac{1}{T}} < \xi < 1$. From this inequality, we have $\eta < \xi^T$. We also have $\xi^T \leq \xi^k$ for all $k = 1, 2, \ldots, T$. Thus we have

$$\frac{\eta}{\xi^k} < 1. \tag{35}$$

The parameter η is selected from the interval $(0.5, 1)$. We also have $\xi^k < 1$. Thus we have

$$0.5 < \eta < \frac{\eta}{\xi^k}. \tag{36}$$

Therefore, the following relationship holds:

$$0.5 < \left(\frac{\eta}{\xi^k} \right)^{\frac{1}{\tau}} < 1. \tag{37}$$

This leads to the following:

$$0.5 < \eta_k' = 1 - \frac{1 - \left(\frac{\eta}{\xi^k} \right)^{\frac{1}{\tau}}}{n_c} < 1. \tag{38}$$

Note that $\boldsymbol{h}_j^\top \boldsymbol{B} \neq \boldsymbol{0}, \forall j = 1, 2, \ldots, n_c$ (Assumption 3), and the left side of (16) is rewritten as follows:

$$\left\| \boldsymbol{h}_j^\top \boldsymbol{B}' \begin{bmatrix} \boldsymbol{\Sigma}_k & \\ & \boldsymbol{\Sigma}_w \end{bmatrix}^{\frac{1}{2}} \right\|_2 = \left\| \left[\boldsymbol{h}_j^\top \boldsymbol{B} \boldsymbol{\Sigma}_k^{\frac{1}{2}},\ \boldsymbol{h}_j^\top \boldsymbol{\Sigma}_w^{\frac{1}{2}} \right] \right\|_{2.} \tag{39}$$

Thus, when

$$\left\| \boldsymbol{h}_j^\top \boldsymbol{\Sigma}_w^{\frac{1}{2}} \right\|_2 \leq \frac{1}{\Phi^{-1}(\eta_k')} (d_j - \boldsymbol{h}_j^\top \hat{\boldsymbol{x}}_{k+1} - \delta_j), \forall \delta_j \in \{\pm \bar{\delta}_j\}, \forall j = 1, \ldots, n_c \tag{40}$$

holds, there exists feasible solutions for (16).

Therefore, from Lemma 1, if the input is determined according to $\boldsymbol{u}_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, the following inequality holds:

$$\Pr\{\boldsymbol{h}_j^\top \boldsymbol{x}_{k+1} \leq d_j\} \geq \eta_k',\ \ \forall j = 1, \ldots, n_c. \tag{41}$$

Note that (41) means that the probability that each state constraint is satisfied is larger than or equal to η_k', while (6) means that the probability that all state constraints are satisfied at the same time is greater than or equal to a certain value. Here, from Bonferroni's inequality we have

$$\Pr\{\boldsymbol{h}_j^\top \boldsymbol{x}_{k+1} \leq d_j,\ \ \forall j = 1, \ldots, n_c\} \geq \sum_{j=1}^{n_c} \Pr\{\boldsymbol{h}_j^\top \boldsymbol{x}_{k+1} \leq d_j\} - (n_c - 1). \tag{42}$$

Therefore,

$$\begin{aligned}
&\Pr\{\boldsymbol{h}_j^\top \boldsymbol{x}_{k+1} \leq d_j\} \geq 1 - \frac{1 - \left(\frac{\eta}{\xi^k}\right)^{\frac{1}{\tau}}}{n_c},\ \ \forall j = 1, \ldots, n_c \\
&\Rightarrow \sum_{j=1}^{n_c} \Pr\{\boldsymbol{h}_j^\top \boldsymbol{x}_{k+1} \leq d_j\} \geq n_c - (1 - \left(\frac{\eta}{\xi^k}\right)^{\frac{1}{\tau}}) \\
&\Leftrightarrow \sum_{j=1}^{n_c} \Pr\{\boldsymbol{h}_j^\top \boldsymbol{x}_{k+1} \leq d_j\} - (n_c - 1) \geq \left(\frac{\eta}{\xi^k}\right)^{\frac{1}{\tau}} \\
&\Rightarrow \Pr\{\boldsymbol{h}_j^\top \boldsymbol{x}_{k+1} \leq d_j,\ \ \forall j = 1, \ldots, n_c\} \geq \left(\frac{\eta}{\xi^k}\right)^{\frac{1}{\tau}} \\
&\Leftrightarrow \Pr\{\boldsymbol{H}\boldsymbol{x}_{k+1} \preceq \boldsymbol{d}\} \geq \left(\frac{\eta}{\xi^k}\right)^{\frac{1}{\tau}}
\end{aligned} \tag{43}$$

holds. That is, (41) is a sufficient condition for

$$\Pr\{\boldsymbol{H}\boldsymbol{x}_{k+1} \preceq \boldsymbol{d}\} \geq \left(\frac{\eta}{\xi^k}\right)^{\frac{1}{\tau}}. \tag{44}$$

Hence, when we determine input by $\boldsymbol{u}_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, chance constraints (6) are satisfied for $k = 1, 2, \ldots, T$.

Next, in the case of (ii) in (14), by determining input as $\boldsymbol{u}_k = \tilde{\boldsymbol{u}}_k$, $\Pr\{\boldsymbol{H}\boldsymbol{x}_{k+1} \preceq \boldsymbol{d}\} \geq \left(\frac{\eta}{\xi^k}\right)^{\frac{1}{\tau}}$ holds from Assumption 5. Therefore, when $\boldsymbol{x}_k \in \mathcal{X}_s$, $\Pr\{\boldsymbol{H}\boldsymbol{x}_{k+1} \preceq \boldsymbol{d}\} \geq \left(\frac{\eta}{\xi^k}\right)^{\frac{1}{\tau}}$ holds by determining input $\boldsymbol{u}_k$ according to (14).

Finally, by determining input as $\boldsymbol{u}_k = \boldsymbol{u}_k^{back}$ in case (iii) of (14), $\Pr\{\boldsymbol{H}\boldsymbol{x}_k \preceq \boldsymbol{d}\} \geq \eta$ holds for any $\boldsymbol{x}_k \in \mathbb{R}^n$ from Lemma 2. Hence, noting $\left(\frac{\eta}{\xi^k}\right)^{\frac{1}{\tau}} > \eta$, $\Pr\{\boldsymbol{H}\boldsymbol{x}_k \preceq \boldsymbol{d}\} \geq \eta$ is satisfied for all time $k = 1, \ldots, T$. $\square$

This theoretical result is obtained with the equivalent transformation of the chance constraints into their deterministic counterparts given in (20) in Lemma 1. This transformation can be used since the disturbance follows a normal distribution. The proposed method would be applicable for dealing with other types of disturbance if at least the sufficient part holds with a certain transformation.

A.4 CONSERVATIVE INPUT THAT DOES NOT INCLUDE EXPLORING ASPECT

Regarding the conservative input that does not contain exploring aspect, we have the following theorem.

Theorem 2. *Let $q \in (0.5, 1)$. Suppose Assumptions 1, 2, 3 and 6 hold. Then, if input $\boldsymbol{u}_k$ satisfies the following inequality for all $j = 1, 2, \dots, n_c$ and $\delta_j \in \{\bar{\delta}_j,\ -\bar{\delta}_j\}$, $\Pr\{\boldsymbol{x}_{k+1} \in \mathcal{X}_s\} \geq q$ holds:*

$$d_j - \boldsymbol{h}_j^\top (\boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{u}_k + \boldsymbol{\mu}_w) - \delta_j - \hat{\delta}_j \geq \Phi^{-1}(q') \left\| \boldsymbol{h}_j^\top \boldsymbol{\Sigma}_w^{\frac{1}{2}} \right\|_2, \tag{45}$$

where $q' = 1 - \frac{1-q}{n_c}$.

In addition, if input sequence $\boldsymbol{U}_k = [\boldsymbol{u}_k^\top, \boldsymbol{u}_{k+1}^\top, \dots, \boldsymbol{u}_{k+\tau-1}^\top]^\top$ satisfies the following inequality for all $j = 1, 2, \dots, n_c$ and $\hat{\delta}_j \in \{-\bar{\hat{\delta}}_j, \bar{\hat{\delta}}_j\}$, $\Pr\{\boldsymbol{x}_{k+\tau} \in \mathcal{X}_s\} \geq q$ holds:

$$d_j - \boldsymbol{h}_j^\top \left(\boldsymbol{A}^\tau \boldsymbol{x}_k + \hat{\boldsymbol{B}}\boldsymbol{U}_k + \hat{\boldsymbol{C}}\hat{\boldsymbol{\mu}}_w \right) - \hat{\delta}_j \geq \Phi^{-1}(q') \left\| \boldsymbol{h}_j^\top \hat{\boldsymbol{C}} \begin{bmatrix} \boldsymbol{\Sigma}_w & & \\ & \ddots & \\ & & \boldsymbol{\Sigma}_w \end{bmatrix}^{\frac{1}{2}} \right\|_2, \tag{46}$$

where $\hat{\boldsymbol{\mu}}_w = \left[\boldsymbol{\mu}_w^\top, \dots, \boldsymbol{\mu}_w^\top\right]^\top \in \mathbb{R}^{n\tau}$, $\hat{\boldsymbol{B}} = [\boldsymbol{A}^{\tau-1}\boldsymbol{B}, \boldsymbol{A}^{\tau-2}\boldsymbol{B}, \dots, \boldsymbol{B}]$ and $\hat{\boldsymbol{C}} = [\boldsymbol{A}^{\tau-1}, \boldsymbol{A}^{\tau-2}, \dots, \boldsymbol{I}]$.

Proof. First, from Bonferroni's inequality, the following relation holds for $q' = 1 - \frac{1-q}{n_c}$:

$$\Pr\{\boldsymbol{H}\boldsymbol{x}_{k+1} \preceq \boldsymbol{d}\} \geq q \Leftarrow \Pr\{\boldsymbol{h}_j^\top \boldsymbol{x}_{k+1} \leq d_j\} \geq q',\ \ \forall j = 1, \dots, n_c. \tag{47}$$

Hence,

$$\begin{aligned} \Pr\{\boldsymbol{H}\boldsymbol{x}_{k+1} \preceq \boldsymbol{d}\} \geq q &\Leftarrow \Pr\left\{\boldsymbol{h}_j^\top (\boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{u}_k + \boldsymbol{e}_k + \boldsymbol{w}_k) \leq d_j\right\} \geq q',\ \forall j = 1, \dots, n_c \\ &\Leftarrow \Pr\left\{\boldsymbol{h}_j^\top (\boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{u}_k + \boldsymbol{w}_k) + \delta_j \leq d_j\right\} \geq q',\ \forall \delta_j,\ \forall j = 1, \dots, n_c, \end{aligned} \tag{48}$$

where $\boldsymbol{e}_k := \boldsymbol{e}(\boldsymbol{x}_k, \boldsymbol{u}_k)$. Next, as in the proof of Lemma 1, the following relation holds:

$$\begin{aligned} &\Pr\{\boldsymbol{h}_j^\top (\boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{u}_k + \boldsymbol{w}_k) + \delta_j \leq d_j\} \geq q' \\ &\Leftrightarrow d_j - \boldsymbol{h}_j^\top (\boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{u}_k) - \delta_j - \boldsymbol{h}_j^\top \boldsymbol{\mu}_w \geq \Phi^{-1}(q') \left\| \boldsymbol{h}_j^\top \boldsymbol{\Sigma}_w^{\frac{1}{2}} \right\|_2. \end{aligned} \tag{49}$$

Therefore, the first part of the theorem is proved.

The state $\boldsymbol{x}_{k+\tau}$ can be expressed by $\boldsymbol{x}_{k+\tau-i}$, $i = 1, 2, \dots, \tau$ as follows:

$$\begin{aligned} \boldsymbol{x}_{k+\tau} =& \boldsymbol{A}\boldsymbol{x}_{k+\tau-1} + \boldsymbol{B}\boldsymbol{u}_{k+\tau-1} + \boldsymbol{e}_{k+\tau-1} + \boldsymbol{w}_{k+\tau-1} \\ =& \boldsymbol{A}\left(\boldsymbol{A}\boldsymbol{x}_{k+\tau-2} + \boldsymbol{B}\boldsymbol{u}_{k+\tau-2} + \boldsymbol{e}_{k+\tau-2} + \boldsymbol{w}_{k+\tau-2}\right) + \boldsymbol{B}\boldsymbol{u}_{k+\tau-1} + \boldsymbol{e}_{k+\tau-1} + \boldsymbol{w}_{k+\tau-1} \\ =& \boldsymbol{A}^2 \boldsymbol{x}_{k+\tau-2} + [\boldsymbol{A}\boldsymbol{B}, \boldsymbol{B}] \left[\boldsymbol{u}_{k+\tau-2}^\top, \boldsymbol{u}_{k+\tau-1}^\top\right]^\top \\ &+ [\boldsymbol{A}, \boldsymbol{I}] \left[\boldsymbol{e}_{k+\tau-2}^\top, \boldsymbol{e}_{k+\tau-1}^\top\right]^\top + [\boldsymbol{A}, \boldsymbol{I}] \left[\boldsymbol{w}_{k+\tau-2}^\top\ \boldsymbol{w}_{k+\tau-1}^\top\right]^\top \\ &\vdots \\ =& \boldsymbol{A}^\tau \boldsymbol{x}_k + \hat{\boldsymbol{B}}\boldsymbol{U}_k + \hat{\boldsymbol{C}}\boldsymbol{E}_k + \hat{\boldsymbol{C}}\boldsymbol{W}_k, \end{aligned} \tag{50}$$

where

$$\begin{aligned} \hat{\boldsymbol{B}} &:= [\boldsymbol{A}^{\tau-1}\boldsymbol{B}, \boldsymbol{A}^{\tau-2}\boldsymbol{B}, \dots, \boldsymbol{B}], \\ \hat{\boldsymbol{C}} &:= [\boldsymbol{A}^{\tau-1}, \boldsymbol{A}^{\tau-2}, \dots, \boldsymbol{I}], \\ \boldsymbol{U}_k &:= [\boldsymbol{u}_k^\top, \boldsymbol{u}_{k+1}^\top, \dots, \boldsymbol{u}_{k+\tau-1}^\top]^\top, \\ \boldsymbol{E}_k &:= [\boldsymbol{e}_k^\top, \boldsymbol{e}_{k+1}^\top, \dots, \boldsymbol{e}_{k+\tau-1}^\top]^\top, \\ \boldsymbol{W}_k &:= [\boldsymbol{w}_k^\top, \boldsymbol{w}_{k+1}^\top, \dots, \boldsymbol{w}_{k+\tau-1}^\top]^\top. \end{aligned}$$

From Bonferroni's inequality, we have

$$\Pr\{\boldsymbol{H}\boldsymbol{x}_{k+\tau} \preceq \boldsymbol{d}\} \geq q \Leftarrow \Pr\{\boldsymbol{h}_j^\top \boldsymbol{x}_{k+\tau} \leq d_j\} \geq q', \ \ \forall j = 1, \ldots, n_c. \tag{51}$$

Next, as in the proof of Lemma 1, the following relation holds:

$$\begin{aligned} &\Pr\{\boldsymbol{h}_j^\top \left(\boldsymbol{A}^\tau \boldsymbol{x}_k + \hat{\boldsymbol{B}}\boldsymbol{U}_k + \hat{\boldsymbol{C}}\boldsymbol{W}_k\right) + \hat{\delta}_j \leq d_j\} \geq q' \\ &\Leftrightarrow d_j - \boldsymbol{h}_j^\top \left(\boldsymbol{A}^\tau \boldsymbol{x}_k + \hat{\boldsymbol{B}}\boldsymbol{U}_k\right) - \hat{\delta}_j - \boldsymbol{h}_j^\top \hat{\boldsymbol{C}}\hat{\boldsymbol{\mu}}_w \geq \Phi^{-1}(q') \left\| \boldsymbol{h}_j^\top \hat{\boldsymbol{C}} \begin{bmatrix} \boldsymbol{\Sigma}_w & & \\ & \ddots & \\ & & \boldsymbol{\Sigma}_w \end{bmatrix}^{\frac{1}{2}} \right\|_2, \end{aligned} \tag{52}$$

where $\hat{\boldsymbol{\mu}}_w = \left[\boldsymbol{\mu}_w^\top, \ldots, \boldsymbol{\mu}_w^\top\right]^\top \in \mathbb{R}^{n \times \tau}$. Therefore, the second part of the theorem is proved. □

Theorem 2 provides sufficient conditions for constructing conservative inputs $\tilde{\boldsymbol{u}}_k$ and $\boldsymbol{u}_k^{back}$.

B Simulation of inverted-pendulum problem

B.1 Simulation conditions

We evaluated the validity of the proposed method with an inverted-pendulum problem provided as "Pendulum-v0" in OpenAI Gym (Brockman et al., 2016). We added external disturbances to this problem and its discrete-time dynamics is given by

$$\begin{bmatrix} \theta_{k+1} \\ \zeta_{k+1} \end{bmatrix} = \begin{bmatrix} \theta_k + T_s \zeta_k \\ \zeta_k - T_s \frac{3g}{2\ell} \sin(\theta_k + \pi) \end{bmatrix} + \begin{bmatrix} 0 \\ T_s \frac{3}{m\ell^2} \end{bmatrix} u_k + \boldsymbol{w}_k, \tag{53}$$

where $\theta_k \in \mathbb{R}$ and $\zeta_k \in \mathbb{R}$ are an angle and angler velocity of the pendulum, respectively. Further, $u_k \in \mathbb{R}$ is an input torque, T_s is a sampling time, and $\boldsymbol{w}_k \in \mathbb{R}^2$ is the external disturbance where

$$\boldsymbol{w}_k \sim \mathcal{N}(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w), \quad \boldsymbol{\mu}_w = \begin{bmatrix} \mu_{w,\theta} \\ \mu_{w,\zeta} \end{bmatrix} \in \mathbb{R}^2, \quad \boldsymbol{\Sigma}_w = \begin{bmatrix} \sigma_{w,\theta}^2 & 0 \\ 0 & \sigma_{w,\zeta}^2 \end{bmatrix} \in \mathbb{R}^{2\times 2}. \tag{54}$$

The details of these and the other variables are shown in Table 1. We let $\boldsymbol{x}_k = [\theta_k, \zeta_k]^\top \in \mathbb{R}^2$ and define a linear approximation model of the above nonlinear system as

$$\boldsymbol{x}_{k+1} \simeq \begin{bmatrix} 1 & T_s \\ 0 & 1 \end{bmatrix} \boldsymbol{x}_k + \begin{bmatrix} 0 \\ T_s \frac{3}{m\ell^2} \end{bmatrix} u_k + \boldsymbol{w}_k. \tag{55}$$

For simplicity, we use the following notations regarding the above system and model:

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} \theta + T_s \zeta \\ \zeta - T_s \frac{3g}{2\ell} \sin(\theta + \pi) \end{bmatrix}, \ \boldsymbol{G} = \begin{bmatrix} 0 \\ T_s \frac{3}{m\ell^2} \end{bmatrix}, \ \boldsymbol{A} = \begin{bmatrix} 1 & T_s \\ 0 & 1 \end{bmatrix}, \ \boldsymbol{B} = \begin{bmatrix} 0 \\ T_s \frac{3}{m\ell^2} \end{bmatrix}.$$

The approximation errors $\boldsymbol{e}$ in (11) is given by

$$\boldsymbol{e}(\boldsymbol{x}, u) = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{G}u - (\boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}u) = \begin{bmatrix} 0 \\ -T_s \frac{3g}{2\ell} \sin(\theta + \pi) \end{bmatrix}. \tag{56}$$

In this verification, we set constraints on ζ_k as $\zeta^{\min} \leq \zeta_k \leq \zeta^{\max}, \forall k = 0, 1, \ldots, T$. This condition becomes

$$\boldsymbol{h}_1^\top \boldsymbol{x}_k \leq d_1, \ \boldsymbol{h}_2^\top \boldsymbol{x}_k \leq d_2, \ \forall k = 0, 1, \ldots, T. \tag{57}$$

where

$$\boldsymbol{h}_1^\top = [0, 1], \ \boldsymbol{h}_2^\top = [0, -1], \ d_1 = \zeta^{\max}, \ d_2 = -\zeta^{\min} \tag{58}$$

and $n_c = 2$. Therefore, Assumption 3 holds since $\boldsymbol{h}_j^\top \boldsymbol{B} \neq 0, \ j \in \{1, 2\}$. Furthermore, the approximation model given in (55) is controllable because of its coefficient matrices $\boldsymbol{A}$ and $\boldsymbol{B}$,

Table 1: Simulation parameters

SYMBOL	DEFINITION	VALUE
T	Number of simulation steps	100
N	Number of learning episodes	100
m	Mass (kg)	1
ℓ	Length of pendulum (m)	1
g	Gravitational const. (m/s^2)	9.8
T_s	Sampling time (s)	0.05
$\boldsymbol{x}_0$	Initial state	$[\pi, 0]^\top$
$\boldsymbol{\mu}_w$	Mean of disturbance	$[0, 0.5]^\top$
$\boldsymbol{\Sigma}_w$	Variance-covariance matrix of disturbance	$\begin{bmatrix} 0.05^2 & 0 \\ 0 & 0.1^2 \end{bmatrix}$
η	Lower bound of probability of constraint satisfaction	0.95
ξ	Lower bound of probability coming back to $\mathcal{X}_s$	0.9998
τ	Maximum steps you need to get back to $\mathcal{X}_s$	2
γ	Discount rate	0.99
α	Learning rates for actor network	1.0×10^{-4}
β	Learning rates for critic network	1.0×10^{-3}

while its controllability index is 2. According to this result, we set $\tau = 2$, and then, $\bar{\delta}_j$ and $\bar{\hat{\delta}}_j$ are given by

$$\bar{\delta}_j = \sup_{\boldsymbol{x} \in \mathbb{R}^2,\, u \in \mathbb{R}} |\boldsymbol{h}_1^\top \boldsymbol{e}(\boldsymbol{x}, u)| = T_s \frac{3g}{2\ell}, \;\; j \in \{1, 2\}, \tag{59}$$

$$\bar{\hat{\delta}}_j = \sup_{\boldsymbol{x} \in \mathbb{R}^2,\, u \in \mathbb{R}} |\boldsymbol{h}_1^\top (\boldsymbol{A} + \boldsymbol{I}) \boldsymbol{e}(\boldsymbol{x}, u)| = T_s \frac{3g}{\ell}, \;\; j \in \{1, 2\}, \tag{60}$$

since $|\sin(\theta + \pi)| \leq 1, \forall \theta \in \mathbb{R}$. As shown in the above equations, the finite upper bounds of the approximation errors are available. We used the bounds given in (59) and (60) in this verification to satisfy Assumption 6.

We have combined our proposed automatic exploration adjustment method (14) with the Deep Deterministic Policy Gradient (DDPG) algorithm (Lillicrap et al., 2015) with immediate cost

$$c_{k+1} = \boldsymbol{\Theta}(\theta_k)^2 + 0.1\theta_k^2 + 0.001u_k^2, \tag{61}$$

where $\boldsymbol{\Theta}(\theta) = \{(\theta + \pi) \mod 2\pi\} - \pi$. Furthermore, in our method, we used the following conservative inputs:

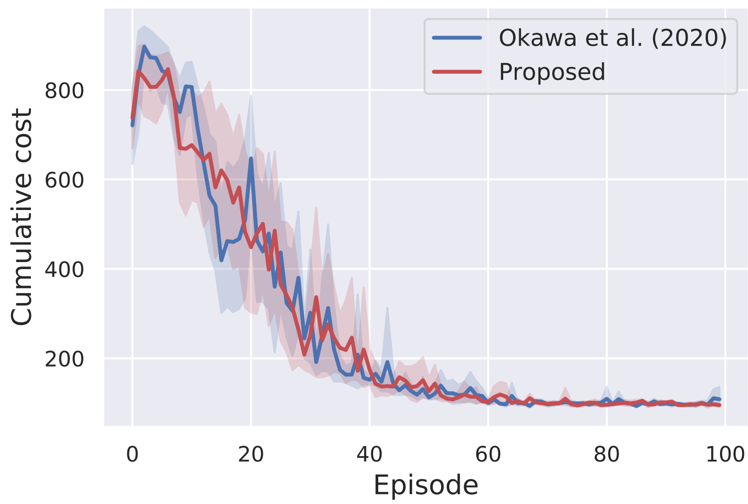
$$\begin{bmatrix} u_k^{back} \\ u_{k+1}^{back} \end{bmatrix} = \begin{bmatrix} -\frac{m\ell^2}{3T_s}(\zeta_k + 2\mu_{w,\theta}) \\ 0 \end{bmatrix}, \; \tilde{u}_k = -\frac{m\ell^2}{3T_s}(\zeta_k + \mu_{w,\theta}). \tag{62}$$

Both of these conservative inputs exist and satisfy the inequalities in Theorem 2 with the parameters listed in Table 1, and thus, Assumptions 4 and 5 are satisfied.

We also combined exploration process adjustment method given in the previous study (Okawa et al., 2020) to DDPG for the reference where $\tilde{u}_k = 0$ as used in that paper. Throughout this verification, we used the same network architectures and hyperparameters as those given in Lillicrap et al. (2015).

B.2 Simulation results

Figure 5 shows the results of the cumulative costs at each episode and the relative frequencies of constraint satisfaction. We evaluated our method and the previous one with 100 episodes $\times$ 10 runs of the simulation (each episode consists of 100 time steps) under the conditions described in Appendix B.1. The results shown in both figures are their mean values, while the shaded areas in the top figure show their 95% confidence intervals. Form these figures, both methods enabled to reduce their cumulative costs as the number of episode increases; however, only the proposed method satisfied the relative frequencies of constraint satisfaction to be equal or greater than η for all steps.



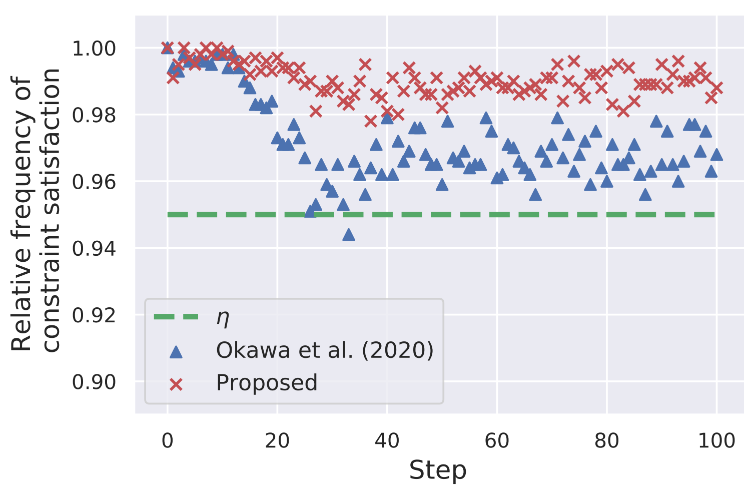


Figure 5: Results on numerical simulation with an inverted-pendulum: (**Top**) Cumulative costs at each episode, (**Bottom**) Relative frequencies of constraint satisfaction at each time step (the same figure as Fig. 5). Both methods enabled to reduce their cumulative costs; however, only the proposed method satisfied the relative frequencies of constraint satisfaction to be equal or greater than η for all steps.