2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Safe Stochastic Model-based Policy Iteration with Chance Constraints

Lijing Zhai[1], Kyriakos G. Vamvoudakis[1], Jérôme Hugues[2]

***Abstract*— In this paper, we consider optimal control problems of stochastic discrete-time systems subject to additive disturbances. Safety of such systems is guaranteed in a probabilistic sense via chance constraints. We solve the corresponding chance constrained stochastic control problems by extending the unconstrained model-based Policy Iteration (PI), and thus chance constrained PI with safety guarantees is proposed. Additionally, the stability of generated control policies is analyzed in the mean square sense. Numerical simulations are provided to validate the proposed algorithm performance.**

I. Introduction

In recent years, there has been growing interest in learning-based control algorithms in academia and industry due to their potential to find optimal control schemes in the presence of uncertainty, where closed-form solutions may be unavailable even with known system dynamics. Safety is essential for learning-based control applied in real-world scenarios. Nevertheless, the trade-off between exploration (learning global optimal policies) and exploitation (finding local optimal policies) can lead to unsafe behavior during learning. To address safety in uncertain or noisy systems, probabilistic chance constraints are preferable to hard constraints, as they ensure safety requirements with high probability, allowing for less conservative control laws. This approach is vital in practical scenarios where some degree of constraint violation is acceptable for economic or optimality reasons. Research in optimal control with chance constraints, including the emerging field of stochastic model predictive control (MPC) [1], has explored various approaches [2]–[8]. Additionally, recent literature has explored the use of control barrier functions (CBFs) for safe control, akin to Lyapunov functions for stability [9]–[11].

Apart from the aforementioned works primarily focusing on addressing chance constraints in the context of optimal control, the rise of learning-based control has led the community to explore safe learning in a probabilistic sense. Two main approaches tackle chance constraints in the learning process. One approach modifies the reward function to explicitly balance risk management and task completion. Frameworks for integrating existing learning algorithms with CBFs are commonly investigated [12]. The other approach focuses on modifying the learning procedure, rather than the reward function, to ensure safe exploration with satisfied constraints. This is achieved by methods like the penalty method (which heavily penalizes constraint violations) and the Lagrangian method (widely used in constrained optimization, with an adaptive weight). A recent work [13] addresses chance-constrained Reinforcement Learning (RL) problems using a combination of the penalty and Lagrangian methods. Most works in this domain employ an actor-critic structure for solving optimization problems by gradient descent methods. The primary challenge lies in approximating gradients of probabilities with respect to parameters [14]. Existing works primarily ensure chance constraint satisfaction but lack theoretical stability guarantees for control policies in systems with stochastic disturbances. Notably, off-policy RL methods may fail to generate stability guarantees in the presence of unknown disturbances [15]. Dealing with both stochastic disturbances and chance constraints adds further complexity to ensuring stability. Our work aims to bridge this gap by investigating optimal control problems in stochastic systems with additive disturbances and chance constraints. We propose a model-based chance-constrained PI method with stability guarantees, which lays the foundation for future data-driven chance-constrained learning algorithms.

Contributions: The contributions of this work are twofold. Safety constraints are embedded in the PI framework in a stochastic setting, and a model-based PI algorithm with chance constraints considered is derived. Additionally, a stability proof of the equilibrium point of the stochastic closed-loop system is provided in the mean square sense.

Notation: $\mathbb{N}$ is the natural number set including zero. $\mathbf{E}$ denotes expectation operator. $\mathbf{P}$ denotes probability operator. $A > 0$ ($A \geq 0$) represents a symmetric positive (non-negative) definite matrix. $\text{tr}(\cdot)$ denotes trace operator. Within the work, matrix norm refers to Frobenius norm while vector norm can be any norms including $\|\cdot\|_1$, $\|\cdot\|_2$, and $\|\cdot\|_\infty$ unless specified. I_n represents an identity matrix with dimension n.

II. Problem Formulation and Preliminaries

A. Problem Description

Consider the following discrete-time linear time-invariant (LTI) dynamical system with additive noise represented by:

$$x_{k+1} = Ax_k + Bu_k + w_k, \tag{1}$$

with discrete time index $k \in \mathbb{N}$, state vector $x_k \in \mathbb{R}^n$, control input $u_k \in \mathbb{R}^m$, state matrix $A \in \mathbb{R}^{n\times n}$, input matrix $B \in \mathbb{R}^{n\times m}$, and exogenous noise/disturbance $w_k \in \mathbb{R}^n$, which can be interpreted as both model uncertainties and

[1]L. Zhai and K. G. Vamvoudakis are with the Daniel Guggenheim School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA, 30332, USA e-mail: lzhai3@gatech.edu, kyriakos@gatech.edu.

[2]J. Hugues is with the Carnegie Mellon University/Software Engineering Institute, Pittsburgh, PA 15213, USA e-mail: jhugues@andrew.cmu.edu.

 This work was supported in part by ONR Minerva under grant No. N00014−18−1−2160, by NSF under grant Nos. CAREER CPS-1851588, CPS-2227185, CPS-2038589, and S&AS 1849198, and by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center. DM23-0307

real noises disturbing the system. Within the work, assume the pair (A, B) is controllable; the disturbance w_k follows an independent and identical Gaussian distribution with a bounded covariance, i.e., $w_k \sim \mathcal{N}(0, \Sigma_{\mathrm{w}})$ with $\|\Sigma_{\mathrm{w}}\| < \infty$; the initial state x_0 is randomly generated from a Gaussian distribution, i.e., $x_0 \sim \mathcal{N}(\mu_0, \Sigma_0)$, and is uncorrelated to w_k.

Assumption 1. Covariance matrices $\Sigma_{\mathrm{w}} \succ 0$, $\Sigma_0 \succ 0$. □

For safety-critical systems in practice, certain safety constraints need to be guaranteed, which are imposed probabilistically as chance constraints on the states as follows:

$$\mathbf{P}\left[h_k^{\mathrm{T}} x_k \geqslant g_k\right] \geqslant 1 - \xi_k, \ \ k \in \mathbb{N}, \tag{2}$$

where $h_k \in \mathbb{R}^n$, $g_k \in \mathbb{R}$, and $\xi_k \in (0, 0.5]$ is a user-defined risk tolerance threshold. Accordingly, define the safe region $\mathcal{S}_k = \{x \in \mathbb{R}^n : \mathbf{P}[h_k^{\mathrm{T}} x_k \geqslant g_k] \geqslant 1 - \xi_k\}, \ \forall k \in \mathbb{N}$. Constraints (2) guarantee that the state trajectories violate the linear constraint $h_k^{\mathrm{T}} x_k \geqslant g_k$ with a probability of at most ξ_k at each time step k. In the stochastic context, stability in the mean square sense is considered. Specifically, *mean square boundedness* is considered [16], [17]:

$$\sup_{k \in \mathbb{N}} \mathbf{E}[\|x_k\|_2^2] < \infty, \ \ \forall x_0 \in \mathcal{S}_0. \tag{3}$$

For a deterministic control policy $\pi : \mathbb{R}^n \to \mathbb{R}^m$ mapping from state x_k to control input u_k, i.e., $u_k = \pi(x_k)$, define the infinite-horizon discounted performance functional:

$$J(x_0, \pi) = \mathbf{E}\left[\sum_{i=0}^{\infty} \gamma^i r(x_i, u_i)\right], \tag{4}$$

where $0 < \gamma < 1$ is a discount factor, $r(x_k, u_k) = x_k^{\mathrm{T}} Q x_k + u_k^{\mathrm{T}} R u_k : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is a stage cost functional incurred at the kth time step and is defined as non-negative with $Q \succeq 0$ and $R \succ 0$. The expectation operation $\mathbf{E}$ is the expectation with respect to the initial state x_0 and the disturbance term w_k. This work aims to determine optimal control policies $\pi^\star(\cdot)$ minimizing the expected discounted cumulative cost (4) with fullfilled chance constraints (2) described by:

Problem 1.

$$\min_{\pi} \ J(x_0, \pi) = \mathbf{E}\left[\sum_{k=0}^{\infty} \gamma^k r(x_k, u_k)\right],$$

$$\text{s.t. } x_{k+1} = A x_k + B u_k + w_k, u_k = \pi(x_k), w_k \sim \mathcal{N}(0, \Sigma_{\mathrm{w}}),$$
$$x_0 \sim \mathcal{N}(\mu_0, \Sigma_0), \mathbf{P}\left[h_k^{\mathrm{T}} x_k \geqslant g_k\right] \geqslant 1 - \xi_k, k \in \mathbb{N}.$$

Definition 1. A control policy $u_k = \pi(x_k)$ is *admissible* for system (1) if the closed-loop system under this policy satisfies (3) and the cost (4) is finite. □

The limit (4) is well-defined for $x_0 \in \mathcal{S}_0$ if the policy π is *admissible* on $\mathcal{S}_k, \ \forall k \in \mathbb{N}$. Assume for system (1) there exist admissible control policies.

B. Stochastic Optimal Control without Chance Constraints

Now we review some existing results on stochastic optimal control problems that do not involve chance constraints (2). These results serve as the foundation for our main results in the next section. Based on the performance functional (4), the cost-to-go value function at time step k is defined as $V(x_k) = \mathbf{E}\left[\sum_{i=k}^{\infty} \gamma^{i-k} r(x_i, u_i)\right]$. The value function can be further rewritten as $V(x_k) = \mathbf{E}[r(x_k, u_k)] + \gamma V(x_{k+1})$. According to Bellman's principle of optimality, the optimal value function $V^\star(\cdot)$ needs to meet the infinite horizon Hamilton-Jacobi-Bellman (HJB) equation [18]:

$$V^\star(x_k) = \min_{\pi(\cdot)} \left(\mathbf{E}[r(x_k, u_k)] + \gamma V^\star(x_{k+1})\right). \tag{5}$$

By stationarity condition for optimality, the optimal control policy $\pi^\star$ should satisfy the first-order necessary condition. This can be achieved by setting the gradient of the right-hand side (RHS) of (5) with respect to u_k equal to 0:

$$\begin{aligned} \pi^\star(x_k) &= \arg\min_{\pi(\cdot)} \left(\mathbf{E}[r(x_k, u_k)] + \gamma V^\star(x_{k+1})\right) \\ &= -\frac{\gamma}{2} R^{-1} B^{\mathrm{T}} \frac{\partial V^\star(x_{k+1})}{\partial x_{k+1}}. \end{aligned} \tag{6}$$

Now consider a linear state feedback control policy $\pi(x_k) = u_k = K x_k$ with $K \in \mathbb{R}^{m \times n}$. The following lemmas provide a sufficient condition for the admissibility of the control policy $u_k = K x_k$ and its corresponding value function.

Lemma 1. *(Lemma 2 in [19]) Assume there is a unique solution $P \in \mathbb{R}^{n \times n}$ to the following Lyapunov equation:*

$$P = \gamma (A + BK)^{\mathrm{T}} P (A + BK) + K^{\mathrm{T}} R K + Q, \tag{7}$$

then the control policy $u_k = K x_k$ is admissible.

Lemma 2. *(Lemma 3 in [19]) Assume the control policy $u_k = K x_k$ is admissible, then the corresponding value function is given by:*

$$V(x_k) = \mathbf{E}[x_k^{\mathrm{T}} P x_k] + \frac{\gamma}{1 - \gamma} tr(P \Sigma_w), \tag{8}$$

where $P \in \mathbb{R}^{n \times n}$ is the unique solution to (7)*.*

Substitute the value function (8) into the RHS of (6):

$$K = -\gamma (R + \gamma B^{\mathrm{T}} P B)^{-1} B^{\mathrm{T}} P A. \tag{9}$$

The Policy Iteration (PI) [20] technique solves the HJB equation iteratively via policy evaluation based on (5) and policy improvement based on (6) with an admissible initial control policy, which is summarized in Algorithm 1 with convergence proof shown by [19].

III. Main Results

Algorithm 1 addresses stochastic optimal control problems without considering chance constraints (2). Inspired by the work [21], our proposed method for solving Problem 1 employs the PI technique to approximate the solution to Problem 1 while incorporating chance constraints during policy improvement. To tackle the challenge posed by chance constraints, we first convert probabilistic constraints into deterministic ones. Then, we establish a chance-constrained PI scheme and finally perform a stability analysis of the resulting control laws in the mean square sense.

Algorithm 1 Policy Iteration without Chance Constraints

1: Select an admissible initial control gain K^0 and threshold η. For $j = 0, 1, \ldots,$ perform until convergence:
2: **repeat**
3: **Policy Evaluation:** Solve for P^j such that

$$P^j = \gamma(A + BK^j)^{\mathrm{T}} P^j (A + BK^j) + (K^j)^{\mathrm{T}} R K^j + Q.$$

4: **Policy Improvement:** Update the policy by

$$K^{j+1} = -\gamma(R + \gamma B^{\mathrm{T}} P^j B)^{-1} B^{\mathrm{T}} P^j A.$$

5: Set $j = j + 1$.
6: **until** $j \geqslant 1$ and $\|P^j - P^{j-1}\| \leqslant \eta$.

A. Representation of Chance Constraints as Deterministic Constraints

Recursively propagating system (1) to get $x_k = A^k x_0 + \sum_{i=0}^{k-1} A^{k-i-1} B u_i + \sum_{i=0}^{k-1} A^{k-i-1} w_i,\ k \geqslant 1$. Under the assumptions of Gaussian distributed x_0 and w_k introduced in Section II-A, the distribution of future states is also Gaussian, i.e., $x_k \sim \mathcal{N}(\mu_k, \Sigma_k)$, with μ_k and Σ_k given by, $\forall k \geqslant 1$:

$$\mu_k = A^k \mu_0 + \sum_{i=0}^{k-1} A^{k-i-1} B u_i, \tag{10}$$

$$\Sigma_k = A^k \Sigma_0 (A^k)^{\mathrm{T}} + \sum_{i=0}^{k-1} A^{k-i-1} \Sigma_{\mathrm{w}} (A^{k-i-1})^{\mathrm{T}}. \tag{11}$$

Remark 1. The state mean at time step k depends on past control inputs up until $k-1$ (i.e., $\{u_0, u_1, \ldots, u_{k-1}\}$) while the state covariance at time step k does not depend on the control inputs. This shows states at future steps are Gaussian distributed with fixed covariance but variant mean. □

Lemma 3. *For a univariate Gaussian random variable $Y \sim \mathcal{N}(\mu, \sigma^2)$ with variant mean μ and fixed variance σ^2, the probabilistic constraint on Y can be converted to a deterministic constraint on its mean as follows:*

$$\mathbf{P}[Y < t] \leqslant \delta \Longleftrightarrow \mu \geqslant c, \tag{12}$$

with $c = t + \sqrt{2}\sigma \cdot \mathit{erf}^{-1}(1 - 2\delta)$,

$$\mathit{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} \mathrm{d}t. \tag{13}$$

Proof. Let $Z = (Y - \mu)/\sigma$. Then $\mathbf{P}[Z < \frac{t-\mu}{\sigma}] = \mathbf{P}[\frac{Y-\mu}{\sigma} < \frac{t-\mu}{\sigma}] = \mathbf{P}[Y < t] \leqslant \delta$. The random variable Z follows a standard Gaussian distribution with mean 0 and variance 1. By inverse standard normal distribution function (also known as inverse error function), we can solve for μ to obtain a deterministic constraint on mean, i.e., $\mu \geqslant t + \sqrt{2}\sigma \cdot \mathrm{erf}^{-1}(1 - 2\delta)$ with error function erf$(\cdot)$ defined by (13) [22]. ■

Based on Lemma 3, the following theorem is introduced to convert the chance constraints (2) to deterministic ones.

Theorem 1. *Assume states follow the propagation rule of* (1)*, then the chance constraints* (2) *on states can be converted to deterministic constraints on their mean as follows:*

$$\mathbf{P}[h_k^{\mathrm{T}} x_k \geqslant g_k] \geqslant 1 - \xi_k \Longleftrightarrow \psi(\mu_k) \geqslant 0, \tag{14}$$

where

$$\psi(\mu_k) = h_k^{\mathrm{T}} \mu_k - g_k - \sqrt{2 h_k^{\mathrm{T}} \Sigma_k h_k} \cdot \mathit{erf}^{-1}(1 - 2\xi_k), \tag{15}$$

with the error function erf$(\cdot)$ defined by (13)*.*

Proof. Given $x_k \sim \mathcal{N}(\mu_k, \Sigma_k)$ for states following the system dynamics (1), the term $h_k^{\mathrm{T}} x_k - g_k$ follows the following Gaussian distribution $h_k^{\mathrm{T}} x_k - g_k \sim \mathcal{N}(h_k^{\mathrm{T}} \mu_k - g_k, h_k^{\mathrm{T}} \Sigma_k h_k)$. Denote $v_k = h_k^{\mathrm{T}} x_k - g_k$, then we have $v_k \sim \mathcal{N}(\mu_v, \sigma_v^2)$ with $\mu_v = h_k^{\mathrm{T}} \mu_k - g_k$ and $\sigma_v = \sqrt{h_k^{\mathrm{T}} \Sigma_k h_k}$. It follows from equation (12) with $t = 0$ in Lemma 3 that $\mathbf{P}[v_k < 0] \leqslant \xi_k \Longleftrightarrow \mu_v \geqslant \sqrt{2}\sigma_v \cdot \mathrm{erf}^{-1}(1 - 2\xi_k)$. Thus, we obtain the deterministic version of the chance constraints $\mathbf{P}[h_k^{\mathrm{T}} x_k \geqslant g_k] \geqslant 1 - \xi_k \Longleftrightarrow \mathbf{P}[h_k^{\mathrm{T}} x_k < g_k] \leqslant \xi_k \Longleftrightarrow h_k^{\mathrm{T}} \mu_k - g_k - \sqrt{2 h_k^{\mathrm{T}} \Sigma_k h_k} \cdot \mathrm{erf}^{-1}(1 - 2\xi_k) \geqslant 0$. ■

B. Chance Constrained Policy Iteration

In order to solve Problem 1 with chance constraints, the optimal value function $V^\star(\cdot)$ needs to meet the HJB equation subject to chance constraints as follows, $k \in \mathbb{N}$:

$$V^\star(x_k) = \min_{\pi(x_k)} \left(\mathbf{E}[r(x_k, \pi(x_k))] + \gamma V^\star(x_{k+1})\right),$$
$$\text{s.t. } x_{k+1} = Ax_k + Bu_k + w_k,\ \mathbf{P}\left[h_k^{\mathrm{T}} x_k \geqslant g_k\right] \geqslant 1 - \xi_k.$$

Regarding the policy evaluation stage in Algorithm 1, the goal is to find the approximate value in terms of the current policy by solving (5). Given an admissible policy π^j, there is always a solution to (5) and thus an approximate solution V^{j+1}. For the policy improvement stage in PI, an updated policy aims to minimize the RHS of (5) with the solved V^{j+1} subject to chance constraints as follows:

$$\begin{aligned} \pi^\star(x_k) = \arg\min_{\pi(\cdot)} &\left(\mathbf{E}[r(x_k, u_k)] + \gamma V^\star(x_{k+1})\right), \\ \text{s.t. } &\mathbf{P}\left[h_k^{\mathrm{T}} x_k \geqslant g_k\right] \geqslant 1 - \xi_k,\ k \in \mathbb{N}. \end{aligned} \tag{16}$$

The chance constraints in equation (16) lack an additive structure, making them challenging to handle. To address this, we apply Theorem 1, which converts chance constraints into deterministic ones. This transformation renders the constrained optimization problem (16) equivalent to

$$\begin{aligned} \pi^\star(x_k) = \arg\min_{\pi(\cdot)} &\left(\mathbf{E}[r(x_k, u_k)] + \gamma V^\star(x_{k+1})\right), \\ \text{s.t. } &\psi(\mu_k) \geqslant 0,\ k \in \mathbb{N}, \end{aligned} \tag{17}$$

with $\psi(\mu_k)$ defined by (15). With this conversion, we can address the chance constraints using the penalty function method. We introduce an additional cost penalty for constraint violation during the policy improvement stage in PI, effectively transforming the constrained optimization problem into an unconstrained one as follows:

$$\begin{aligned} \pi^\star(x_k) = \arg\min_{\pi(\cdot)} \Big(&\mathbf{E}[r(x_k, u_k)] + \gamma V^\star(x_{k+1}) \\ &+ \lambda \max(-\psi(\mu_{k+1}),\ 0)^2\Big), \end{aligned} \tag{18}$$

where $\lambda > 0$ is a penalization parameter. This formulation implies that in the policy improvement stage at time step k, once the chance constraint at time step $k+1$ under current

control input is violated, i.e., $\psi(\mu_{k+1}) < 0$, the penalized cost is enforced to generate a feasible policy within the safe region such that the constraint at time step $k+1$ is satisfied.

Remark 2. We draw inspiration from recent work [21], which deals with a continuous-time system subject to hard constraints. Here, in contrast, we incorporate chance constraints for a single future time step in the infinite horizon setting, while the approach in [21] integrates constraints from the current time instant to the final finite time horizon. □

With the integration of chance constraints in the policy improvement stage by the penalty function method, the following theorem is introduced to derive the solution to the unconstrained optimization problem (18).

Theorem 2. *The solution to optimization problem* (18) *is given by,* $\forall k \geqslant 1$*:*

$$\pi^\star(x_k) = u_k^\star = \begin{cases} Kx_k, & \text{if } \psi(\mu_{k+1}) \geqslant 0 \\ K_c x_k + u_c, & \text{otherwise} \end{cases} \tag{19}$$

where

$$K = -\gamma(R + \gamma B^{\mathrm{T}} P B)^{-1} B^{\mathrm{T}} P A, \tag{20}$$

$$K_c = -\gamma(R + \gamma B^{\mathrm{T}} P B + \lambda B^{\mathrm{T}} h_{k+1} h_{k+1}^{\mathrm{T}} B)^{-1} B^{\mathrm{T}} P A, \tag{21}$$

$$u_c = -\lambda(R + \gamma B^{\mathrm{T}} P B + \lambda B^{\mathrm{T}} h_{k+1} h_{k+1}^{\mathrm{T}} B)^{-1} \varphi_{k+1}, \tag{22}$$

$$\varphi_{k+1} = \Bigg(h_{k+1}^{\mathrm{T}} A^{k+1} \mu_0 + h_{k+1}^{\mathrm{T}} \sum_{i=0}^{k-1} A^{k-i} B u_i - g_{k+1} - \sqrt{2 h_{k+1}^{\mathrm{T}} \Sigma_{k+1} h_{k+1}} \cdot \mathit{erf}^{-1}(1 - 2\xi_{k+1}) \Bigg) B^{\mathrm{T}} h_{k+1}. \tag{23}$$

Proof. To solve the unconstrained optimization problem (18), we apply the first-order necessary condition for optimality by setting the derivative of RHS of equation (18) with respect to u_k equal to 0. If $\psi(\mu_{k+1}) \geqslant 0$, there is no extra penalty cost and thus the optimal control policy is the same to (6) with the state feedback gain given by (9). If $\psi(\mu_{k+1}) < 0$, the penalty cost is enforced. Given value function (8), state mean (10) and deterministic constraints (15), set the derivative of RHS of equation (18) with respect to u_k to zero to obtain: $0 = 2Ru_k + \gamma B^{\mathrm{T}} \frac{\partial V^\star(x_{k+1})}{\partial x_{k+1}} + 2\lambda \psi(\mu_{k+1}) \frac{\partial \psi(\mu_{k+1})}{\partial u_k} = 2Ru_k + 2\gamma B^{\mathrm{T}} P B u_k + 2\lambda h_{k+1}^{\mathrm{T}} B u_k B^{\mathrm{T}} h_{k+1} + 2\gamma B^{\mathrm{T}} P A x_k + 2\lambda \varphi_{k+1}$. Note that the term $\psi(\mu_{k+1})$ contains u_k due to μ_{k+1}, whereas the term $\varphi(\mu_{k+1})$ does not contain u_k. So in the above manipulation, we substitute the state mean (10) to separate u_k from $\psi(\mu_{k+1})$. The objective is to rearrange the terms such that those containing u_k are on one side with all remaining terms on the other side. Note that the term $h_{k+1}^{\mathrm{T}} B u_k B^{\mathrm{T}} h_{k+1}$ contains u_k in the middle. Applying trace operation to both sides and leveraging the cyclic property of trace operation: $\mathrm{tr}\left((R + \gamma B^{\mathrm{T}} P B + \lambda B^{\mathrm{T}} h_{k+1} h_{k+1}^{\mathrm{T}} B) u_k\right) = \mathrm{tr}(-\gamma B^{\mathrm{T}} P A x_k - \lambda \varphi_{k+1})$. The summation term $R + \gamma B^{\mathrm{T}} P B + \lambda B^{\mathrm{T}} h_{k+1} h_{k+1}^{\mathrm{T}} B \succ 0$ since $R \succ 0$, $P \succ 0$, and $B^{\mathrm{T}} h_{k+1} h_{k+1}^{\mathrm{T}} B \succeq 0$, and thus is invertible. So,

$$\begin{aligned} u_k^\star = & -(R + \gamma B^{\mathrm{T}} P B + \lambda B^{\mathrm{T}} h_{k+1} h_{k+1}^{\mathrm{T}} B)^{-1} (\gamma B^{\mathrm{T}} P A x_k \\ & + \lambda \varphi_{k+1}) \\ = & \underbrace{-\gamma(R + \gamma B^{\mathrm{T}} P B + \lambda B^{\mathrm{T}} h_{k+1} h_{k+1}^{\mathrm{T}} B)^{-1} B^{\mathrm{T}} P A}_{K_c} x_k \\ & \underbrace{-\lambda(R + \gamma B^{\mathrm{T}} P B + \lambda B^{\mathrm{T}} h_{k+1} h_{k+1}^{\mathrm{T}} B)^{-1} \varphi_{k+1}}_{u_c}. \end{aligned}$$

This completes the proof. ■

Remark 3. For the case of $\psi(\mu_{k+1}) < 0$, the extra term u_c acts as a compensation controller to compensate for the unknown disturbance and violation of constraints. □

Based on Theorem 2, PI with chance constraints is proposed as Algorithm 2, using policy evaluation based on (5) and policy improvement based on (18). Algorithm (18) contains two steps. First, the offline PI is implemented to obtain the converged P and K. Then, for each time step, assume control policies u_k with control gain K. If the constraint (23) is violated, compute the controller gain K_c and the compensation controller u_c, and apply $u_k = K_c + u_c$ to system (1) to compensate for constraint violation.

Algorithm 2 Policy Iteration with Chance Constraints

1: Select an admissible initial control gain K^0, convergence threshold η, threshold ξ, penalty parameter λ, and time steps T. For $j = 0, 1, \ldots$, perform until convergence:
2: **repeat**
3: Solve for P^j such that
$$P^j = \gamma(A + BK^j)^{\mathrm{T}} P^j (A + BK^j) + (K^j)^{\mathrm{T}} R K^j + Q.$$
4: Update the policy by
$$K^{j+1} = -\gamma(R + \gamma B^{\mathrm{T}} P^j B)^{-1} B^{\mathrm{T}} P^j A.$$
5: Set $j = j + 1$.
6: **until** $j \geqslant 1$ and $\|P^j - P^{j-1}\| \leqslant \eta$.
7: **for** $k = 0$ to T **do**
8: Set $u_k = K^{j+1} x_k$.
9: **if** $\psi(\mu_{k+1}) < 0$ **then**
10: $K_c = -\gamma(R + \gamma B^{\mathrm{T}} P B + \lambda B^{\mathrm{T}} h_{k+1} h_{k+1}^{\mathrm{T}} B)^{-1} \times B^{\mathrm{T}} P A$
11: set $u_k = K_c + u_c$ with u_c given by (22).
12: **end if**
13: Apply u_k to system (1) and store the next state x_{k+1}.
14: **end for**

C. Stability Analysis

This section will focus on analyzing the stability of generated control policies by chance constrained PI Algorithm 2. We will demonstrate the boundedness of the control policies (19) and then prove the mean square boundedness for the closed-loop system (1) under these policies.

Lemma 4. *The optimal control policies* (19) *are bounded, i.e.,* $\|u_k\|_\infty \leqslant U_{max}$, $\forall k \in \mathbb{N}$.

Proof. When $\varphi(\mu_{k+1}) \geqslant 0$, the control gain K given by (20) is the same to that by Algorithm 1 and thus is bounded. When

$\varphi(\mu_{k+1}) < 0$, consider the terms R and $\lambda B^{\mathrm{T}} h_{k+1} h_{k+1}^{\mathrm{T}} B$ in (21) together as one single term so that K_{c} has the same structure to K given by (20) and thus is bounded. u_{c} is bounded since ϕ_{k+1} contains a summation of finite bounded terms. So, u_k is bounded in both cases. ∎

With Lemma 4, the stability of closed-loop systems under the control policies (19) in the mean square sense is verified.

Theorem 3. *Assume the matrix A is Schur stable. System* (1) *with chance constraints* (2) *under the control policies* (19) *is mean square bounded.*

Proof. The proof follows a similar reasoning to that in [23]. For system (1), we have $\mathbf{E}[x_{k+1}^{\mathrm{T}} P x_{k+1}] = \mathbf{E}[(Ax_k + Bu_k + w_k)^{\mathrm{T}} P (Ax_k + Bu_k + w_k)] = \mathbf{E}[x_k^{\mathrm{T}} A^{\mathrm{T}} P A x_k + 2x_k^{\mathrm{T}} A^{\mathrm{T}} P B u_k + 2x_k^{\mathrm{T}} A^{\mathrm{T}} P w_k + u_k^{\mathrm{T}} B^{\mathrm{T}} P B u_k + 2u_k^{\mathrm{T}} B^{\mathrm{T}} P w_k + w_k^{\mathrm{T}} P w_k] = x_k^{\mathrm{T}} A^{\mathrm{T}} P A x_k + 2x_k^{\mathrm{T}} A^{\mathrm{T}} P B u_k + u_k^{\mathrm{T}} B^{\mathrm{T}} P B u_k + \mathrm{tr}(P\Sigma_{\mathrm{w}})$. Considering the boundedness of u_k and using Hölder's inequality, i.e., $\|x^{\mathrm{T}} y\|_1 \leqslant \|x\|_p \|y\|_q$ with $\frac{1}{p} + \frac{1}{q} = 1$, the terms on the RHS of the above equation are bounded by $\|x_k^{\mathrm{T}} A^{\mathrm{T}} P B u_k\|_1 \leqslant \|A^{\mathrm{T}} P B u_k\|_1 \|x_k\|_\infty \leqslant \|A^{\mathrm{T}} P B\|_1 \|u_k\|_\infty \|x_k\|_\infty \leqslant \|A^{\mathrm{T}} P B\|_1 U_{\max} \|x_k\|_\infty$ and $\|u_k^{\mathrm{T}} B^{\mathrm{T}} P B u_k\|_1 \leqslant \|B^{\mathrm{T}} P B\|_1 U_{\max}^2$. Then it follows that $\mathbf{E}[x_{k+1}^{\mathrm{T}} P x_{k+1}] \leqslant x_k^{\mathrm{T}} A^{\mathrm{T}} P A x_k + 2c_1 \|x_k\|_\infty + c_2$ with $c_1 = \|A^{\mathrm{T}} P B\|_1 U_{\max}$, $c_2 = \|B^{\mathrm{T}} P B\|_1 U_{\max}^2 + \mathrm{tr}(P\Sigma_{\mathrm{w}})$. Given $A^{\mathrm{T}} P A - P \leqslant -I_n$ and Schur stable matrix A,

$$\mathbf{E}[x_{k+1}^{\mathrm{T}} P x_{k+1}] \leqslant x_k^{\mathrm{T}} P x_k - \|x_k\|_2^2 + 2c_1 \|x_k\|_\infty + c_2. \tag{24}$$

Define a compact set $\mathcal{D} = \{x \in \mathbb{R}^n : \|x_k\|_\infty \leqslant \beta\}$ with $\beta = \frac{1}{\theta}(c_1 + \sqrt{c_1^2 + c_2\theta})$ and $\theta \in (\max\{0, 1 - \lambda_{\max}(P)\}, 1)$. Note that the matrix A being Schur stable guarantees that such θ exists. Then it follows that

$$2c_1 \|x_k\|_\infty + c_2 \leqslant \theta \|x_k\|_\infty^2 \leqslant \theta \|x_k\|_2^2, \ \forall x_k \notin \mathcal{D}. \tag{25}$$

Substitute (25) into (24) to get $\mathbf{E}[x_{k+1}^{\mathrm{T}} P x_{k+1}] \leqslant x_k^{\mathrm{T}} P x_k - (1-\theta)\|x_k\|_2^2, \ \forall x_k \notin \mathcal{D}$. Combine the above equation with $x_k^{\mathrm{T}} P x_k \leqslant \lambda_{\max}(P) \|x_k\|_2^2$, then we get $\mathbf{E}[x_{k+1}^{\mathrm{T}} P x_{k+1}] \leqslant (1 - \frac{1-\theta}{\lambda_{\max}}) x_k^{\mathrm{T}} P x_k, \ \forall x_k \notin \mathcal{D}$. This leads to $\sup_{k\in\mathbb{N}} V(x_k) < \infty$ according to Lemma 8 in [23]. Then it follows from $\lambda_{\min}(P)\|x_k\|_2^2 \leqslant x_k^{\mathrm{T}} P x_k$ and the value function (8) that $\sup_{k\in\mathbb{N}} \mathbf{E}[\|x_k\|_2^2] \leqslant \frac{1}{\lambda_{\min}(P)} \mathbf{E}[x_k^{\mathrm{T}} P x_k] = \frac{1}{\lambda_{\min}(P)} \left(V(x_k) - \frac{\gamma}{1-\gamma} \mathrm{tr}(P\Sigma_{\mathrm{w}}) \right) < \infty$. ∎

IV. Simulation Results

Consider a quadrotor described by [24], $\forall k$,

$$A = \begin{bmatrix} 0.99 & 0 & 0 & 0.02 & 0 & 0 \\ 0 & 0.99 & 0 & 0 & 0.02 & 0 \\ 0 & 0 & 0.99 & 0 & 0 & 0.02 \\ 0 & 0 & -0.02 & 0.99 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.99 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.99 \end{bmatrix},$$

$$B = \begin{bmatrix} 0 & 0 & 0.06 & 0 & 0.02 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.02 \end{bmatrix}^{\mathrm{T}},$$

where the states are $x_k = [x, \ \dot{x}, \ y, \ \dot{y}, \ \phi, \ \dot{\phi}]^{\mathrm{T}}$ with (x, y) the two-dimensional quadrotor positions, ϕ the counter-clockwise angle to the vertical, u_1 is the vertical thrust, and u_2 is the torque. We slightly modify matrix A to ensure Schur stability, as per the assumption in this work compared to that in [24]. The safety constraint enforces $x_3 \geqslant 0.05$ (y position), with $h_k = [0, \ 0, \ 1, \ 0, \ 0, \ 0]^{\mathrm{T}}$ and $g_k = 0.05$. Disturbance covariance is generated as random numbers from a normal distribution in the range of [0,1], scaled by 0.001. Weight matrices are set as $R = 1$ and $Q = 10 I_2$. The system is simulated for 40 seconds with a sampling time of 0.1 seconds, resulting in a total of 400 time steps.

As a benchmark, we first employ model-based PI without chance constraints using Algorithm 1 in 500 independent implementations. Figure 1 displays state trajectories, with the blue line representing the mean, the shaded area indicating a 75% confidence interval, and the red dotted line representing the safety constraint $x_3 = 0.05$. Algorithm 1 effectively stabilizes the system with states' mean values close to zero. However, due to stochastic disturbances, the 75% confidence interval for x_3 is wide, indicating multiple violations of the safety constraint $x_3 \geqslant 0.05$. This implies that not considering chance constraints (2) increases the likelihood of violating safety constraints for state x_3.

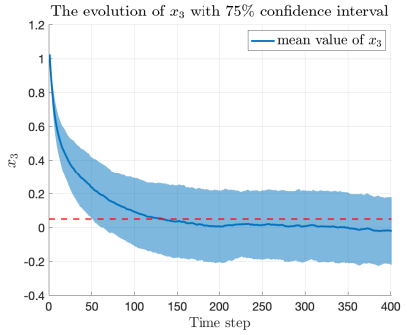
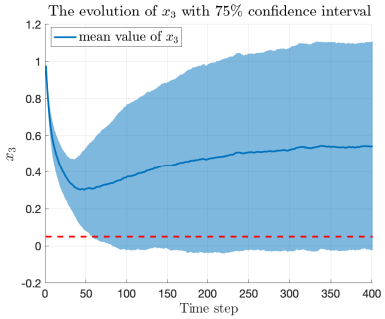


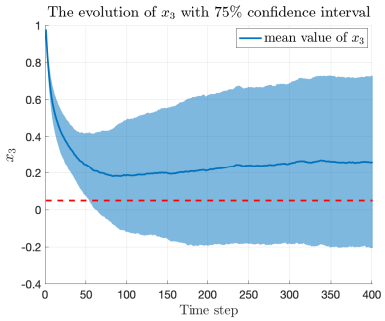
Fig. 1. State trajectories by the PI Algorithm 1 without chance constraints. The shaded area represents a 75% confidence interval for 500 independent experiments. The red dotted line denotes the safety constraint of $x_3 = 0.05$.

Next, we explore different risk tolerance levels, namely $\xi = 0.2$ and $\xi = 0.45$, within the same simulation framework using Algorithm 2, conducting 500 independent experiments. Figure 2 illustrates state trajectories, with the blue line denoting the mean, shaded areas indicating 75% confidence intervals, and the red dotted line representing the safety constraint $x_3 = 0.05$. Among the 500 experiments, a certain number of state trajectories violate safety constraints in both cases, as expected due to probabilistic constraints. However, the mean state value with a risk tolerance threshold of $\xi = 0.2$ deviates further from safety constraints compared to $\xi = 0.45$. This suggests that smaller risk thresholds result in fewer trajectories breaching safety constraints.

Our proposed algorithm also accommodates time-varying chance constraints. Consider time-varying chance constraints defined by $h_k = [0, \ 0, \ 1, \ 0, \ 0, \ 0]^{\mathrm{T}}$ and $g_k = 0.05 + v_k$, where v_k is a randomly generated number from a uniform distribution between 0 and 0.05. Figure 3 displays x_3 trajectories, indicating that the control policies generated by Algorithm 2 effectively ensure a high probability of fullfillment with the time-varying safety constraints.



(a) Risk tolerance threshold $\xi = 0.2$



(b) Risk tolerance threshold $\xi = 0.45$

Fig. 2. State trajectories by Algorithm 2 with chance constraints. Shaded areas indicate 75% confidence intervals. The red dotted line denotes $x_3 = 0.05$. Smaller thresholds lead to fewer trajectories violating constraints.

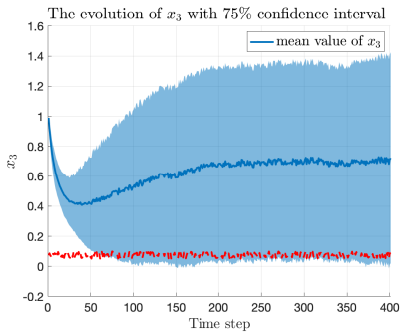


Fig. 3. Trajectories of x_3 by PI Algorithm 2 with chance constraints. The shaded area represents a 75% confidence interval for 500 independent experiments. The red dotted line denotes the time-varying safety constraints.

V. Conclusion and Future Work

This work addresses optimal control problems in stochastic discrete-time systems with additive disturbances and chance constraints. We introduce a model-based PI algorithm with chance constraints and analyze the stability of control policies in the mean square boundedness sense. We validate the algorithm's performance through numerical simulations for a steering system in autonomous vehicles. Future work will explore the relationship between risk tolerance (ξ) and penalty parameter (λ), handle joint chance constraints, and extend the algorithm to data-based learning frameworks.

References

[1] M. Farina, L. Giulioni, and R. Scattolini, "Stochastic linear model predictive control with chance constraints–a review," *Journal of Process Control*, vol. 44, pp. 53–67, 2016.

[2] G. Schildbach, P. Goulart, and M. Morari, "The linear quadratic regulator with chance constraints," in *2013 European Control Conference (ECC)*. IEEE, 2013, pp. 2746–2751.

[3] S. Nandi and T. Singh, "Chance constraint based design of controllers for linear uncertain systems," in *2017 American Control Conference (ACC)*. IEEE, 2017, pp. 4510–4515.

[4] S. Paternain, M. Calvo-Fullana, L. F. Chamon, and A. Ribeiro, "Learning safe policies via primal-dual methods," in *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 2019, pp. 6491–6497.

[5] J. Pilipovsky and P. Tsiotras, "Chance-constrained optimal covariance steering with iterative risk allocation," in *2021 American Control Conference (ACC)*. IEEE, 2021, pp. 2011–2016.

[6] P. Hokayem, D. Chatterjee, and J. Lygeros, "Chance-constrained lqg with bounded control policies," in *52nd IEEE Conference on Decision and Control*. IEEE, 2013, pp. 2471–2476.

[7] M. Ono, M. Pavone, Y. Kuwata, and J. Balaram, "Chance-constrained dynamic programming with application to risk-aware robotic space exploration," *Autonomous Robots*, vol. 39, pp. 555–571, 2015.

[8] H. Zhong, Y. Shimizu, and J. Chen, "Chance-constrained iterative linear-quadratic stochastic games," *IEEE Robotics and Automation Letters*, vol. 8, no. 1, pp. 440–447, 2022.

[9] A. Clark, "Control barrier functions for complete and incomplete information stochastic systems," in *2019 American Control Conference (ACC)*. IEEE, 2019, pp. 2928–2935.

[10] F. B. Mathiesen, S. C. Calvert, and L. Laurenti, "Safety certification for stochastic systems via neural barrier functions," *IEEE Control Systems Letters*, vol. 7, pp. 973–978, 2022.

[11] J. Xu, J. Wang, J. Rao, Y. Zhong, and H. Wang, "Adaptive dynamic programming for optimal control of discrete-time nonlinear system with state constraints based on control barrier function," *International Journal of Robust and Nonlinear Control*, vol. 32, no. 6, pp. 3408–3424, 2022.

[12] Z. Marvi and B. Kiumarsi, "Safe off-policy reinforcement learning using barrier functions," in *2020 American Control Conference (ACC)*. IEEE, 2020, pp. 2176–2181.

[13] B. Peng, J. Duan, J. Chen, S. E. Li, G. Xie, C. Zhang, Y. Guan, Y. Mu, and E. Sun, "Model-based chance-constrained reinforcement learning via separated proportional-integral lagrangian," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[14] A. Geletu, A. Hoffmann, and P. Li, "Analytic approximation and differentiability of joint chance constraints," *Optimization*, vol. 68, no. 10, pp. 1985–2023, 2019.

[15] R. Song, F. L. Lewis, Q. Wei, and H. Zhang, "Off-policy actor-critic structure for optimal control of unknown systems with disturbances," *IEEE transactions on cybernetics*, vol. 46, no. 5, pp. 1041–1050, 2015.

[16] D. Chatterjee, F. Ramponi, P. Hokayem, and J. Lygeros, "On mean square boundedness of stochastic linear systems with bounded controls," *Systems & Control Letters*, vol. 61, no. 2, pp. 375–380, 2012.

[17] J. A. Paulson, E. A. Buehler, R. D. Braatz, and A. Mesbah, "Stochastic model predictive control with joint chance constraints," *International Journal of Control*, vol. 93, no. 1, pp. 126–139, 2020.

[18] B. O'Donoghue, Y. Wang, and S. Boyd, "Iterated approximate value functions," in *2013 European Control Conference (ECC)*. IEEE, 2013, pp. 3882–3888.

[19] J. Lai, J. Xiong, and Z. Shu, "Model-free optimal control of discrete-time systems with additive and multiplicative noises," *Automatica*, vol. 147, p. 110685, 2023.

[20] R. S. Sutton and A. G. Barto, "Reinforcement learning: an introduction mit press," *Cambridge, MA*, vol. 22447, 1998.

[21] Z. Lin, J. Ma, J. Duan, S. E. Li, H. Ma, B. Cheng, and T. H. Lee, "Policy iteration based approximate dynamic programming toward autonomous driving in constrained dynamic environment," *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[22] L. Blackmore, H. Li, and B. Williams, "A probabilistic approach to optimal robust path planning with obstacles," in *2006 American Control Conference*. IEEE, 2006, pp. 7–pp.

[23] P. Hokayem, D. Chatterjee, and J. Lygeros, "On stochastic model predictive control with bounded control inputs," *arXiv preprint arXiv:0902.3944*, 2009.

[24] S. Pfrommer, T. Gautam, A. Zhou, and S. Sojoudi, "Safe reinforcement learning with chance-constrained model predictive control," in *Learning for Dynamics and Control Conference*. PMLR, 2022, pp. 291–303.