arXiv:2103.01413v1 [cs.LG] 2 Mar 2021

Safe Learning of Uncertain Environments for Nonlinear Control-Affine Systems

Farhad Farokhi*, Alex Leong†, Iman Shames‡, and Mohammad Zamani§

Abstract

In many learning based control methodologies, learning the unknown dynamic model precedes the control phase, while the aim is to control the system such that it remains in some safe region of the state space. In this work our aim is to guarantee safety while learning and control proceed simultaneously. Specifically, we consider the problem of safe learning in nonlinear control-affine systems subject to unknown additive uncertainty. We model uncertainty as a Gaussian signal and use state measurements to learn its mean and covariance. We provide rigorous time-varying bounds on the mean and covariance of the uncertainty and employ them to modify the control input via an optimisation program with safety constraints encoded as a barrier function on the state space. We show that with an arbitrarily large probability we can guarantee that the state will remain in the safe set, while learning and control are carried out simultaneously, provided that a feasible solution exists for the optimisation problem. We provide a secondary formulation of this optimisation that is computationally more efficient. This is based on tightening the safety constraints to counter the uncertainty about the learned mean and covariance. The magnitude of the tightening can be decreased as our confidence in the learned mean and covariance increases (i.e., as we gather more measurements about the environment). Extensions of the method are provided for Gaussian uncertainties with piecewise constant mean and covariance to accommodate more general environments.

1 Introduction

1.1 Motivation

Safety of a dynamically controlled system can be defined as guaranteeing that the closed loop system trajectory remains inside a subset of its state space, denoted by the safety set. The Control Barrier Function (CBF) approach and many control strategies achieves this by requiring a reliable dynamic model of the system. However, dynamical models are typically subject to uncertainties. There are many approaches to deal with these uncertainties. Namely, one can model uncertainty as deterministic or stochastic signals with known (upper bound) size or statistics. In these cases, the control design aims to guarantee safety for the worst-case realisation of the uncertainty. In more recent literature, the attempt is to learn the uncertain parts of the dynamics from historic data of the system in order to obtain more accurate and less conservative models of the uncertain parts of the system model. In many such approaches a commonality is that learning the uncertain parameters is carried out separately (prior to control), to ensure an accurate model of the uncertainties is available before attempting to control the system safely. In this paper, we weaken this assumption by learning the uncertainties while maintaining safety through appropriate manipulations of the control signal.

1.2 Contributions

In this paper, we propose a learning based control methodology that guarantees safety while both learning and control are carried out simultaneously. In our approach, we consider a nonlinear control-affine model

*The University of Melbourne. e-mail: ffarokhi@unimelb.edu.au
†Defence Science and Technology Group, e-mail: alex.leong@dst.defence.gov.au
‡Australian National University, e-mail: iman.shames@anu.edu.au
§Defence Science and Technology Group, e-mail: mohammad.zamani@dst.defence.gov.au

subject to a Gaussian process noise with unknown parameters. Safety of the agent is characterised by time-varying state constraints which at each time step bars the agent from parts of its state space.

We start by considering a zero-mean Gaussian noise uncertainty with an unknown covariance matrix, which is learnt online from state measurements. **Lemma 2** shows that with probability greater than $1-\delta$, the actual covariance is upper bounded by its empirical estimated covariance plus a positive offset term. The offset term is bigger for more stringent guarantees (smaller δ), but more importantly shrinks as time grows (more historic data is available for learning the covariance estimate).

Our approach to safe controller design is via an optimisation problem (4) that seeks the closest control signal with respect to a nominal control input, subject to state constraints characterised with respect to the bounds derived in Lemma 2. Our first main result, **Theorem 3**, shows that solving this optimisation problem, provided that it is feasible, will ensure safety with probability greater than $1-\delta$ at each time.

Note that Theorem 3 allows us to balance the required degree of safety guarantee, depending on whether there is sufficient data and whether the optimisation problem (4) is feasible. However, this optimisation problem is potentially difficult to solve numerically as it has infinitely many constraints (for each noise parameter characterised in Lemma 2). We address this problem in **Lemma 4**, where it is shown that the infinite number of constraints is equivalent to a single constraint characterised using the bound on the noise. This helps us reformulate our optimisation problem as (5), a problem with a safety constraint involving a tightening term. This term is also bigger for more stringent safety guarantee requirements, but shrinks with time as more data becomes available for covariance estimation.

We then provide a couple of extensions for the aforementioned results. Namely, we first tackle the non-zero-mean Gaussian noise case in Section 3. In this case, both the mean and covariance of the noise are learned from state measurements. **Lemma 6** and **Lemma 7** provide bounds on the mean and the covariance of the noise based on their empirically estimated values. Safe control of a system subject to a non-zero-mean Gaussian uncertainty is then formulated as optimisation problem (9). **Theorem 8** (counterpart to Theorem 3) then states that if this optimisation is feasible, then the control signal obtained can guarantee safety of the next visited state with probability greater than $1-\delta$.

Finally, in Section 4 we address the challenging scenario where the noise signal does not have fixed parameters but is rather state dependent. This scenario is relevant for many robotic applications where the environment is not homogeneous (e.g. has different terrains), or many control scenarios where the un-modelled dynamics of the agent is different based on the part of the state-space being explored. We model these scenarios by noise signals that have mean and covariance being piece-wise constant in different partitions of the state space. We then proceed to estimate these parameters for each partition based on state data obtained in those areas. In **Corollary 9** we provide bounds on the estimated mean and covariance for each partition, similar to **Lemma 6** and **Lemma 7**. Assuming that there is some regularity, i.e., these parameters do not change abruptly across neighbouring partitions (**Assumption 10**), we derive alternative bounds in **Lemma 12** and **Lemma 13** which in some cases are tighter than Corollary 9.

1.3 Related Work

Early works on using control barrier functions for safety of control systems include (Ames et al., 2014, 2017). A notion of input-to-state safe control barrier functions was introduced in (Kolathaya and Ames, 2019). These works however require accurate models of the dynamical system.

For nonlinear control-affine systems, safe control using control barrier functions, for systems where model uncertainties are unknown and learnt include (Cheng et al., 2019, 2020; Choi et al., 2020; Jagtap et al., 2020; Taylor et al., 2020). The works (Choi et al., 2020; Taylor et al., 2020) attempt to learn the form of the uncertainty using data-driven approaches, but no rigorous guarantees on safety are provided. The model uncertainties are modelled using Gaussian processes in (Cheng et al., 2019, 2020; Jagtap et al., 2020), with the aim of learning the parameters of these Gaussian processes to then derive safety bounds. All of these works assume either that the model uncertainty (Choi et al., 2020; Jagtap et al., 2020) or various model parameters of the Gaussian process (Cheng et al., 2020) are learnt before control, or learnt in batch mode (Cheng et al., 2019; Taylor et al., 2020) while alternating with controller synthesis. The current paper differs from these works in that both learning and control are done simultaneously after every new state measurement, while providing rigorous theoretical bounds that the system will be safe with high probability at all times.

Other approaches have also been proposed in the area of safe reinforcement learning (Garcia and Fernandez, 2015), where control signals/actions are learnt using reinforcement learning techniques, but modified to ensure that the system is safe (Junges et al., 2016; Yu et al., 2019; Zhang et al., 2019). These modifications include constraining exploration to strategies which are safe (Junges et al., 2016), solving a constrained Markov decision process where the constraints provide safety (Yu et al., 2019), and the use of shielding (Zhang et al., 2019) with a backup controller which is guaranteed to be safe. The algorithms employed usually learn the actions in a "model-free" manner, even if some knowledge of the system may be available, and as such the performance may be conservative. Furthermore, without a model, safety often can not be guaranteed during learning but only after a sufficiently long learning period (Cheng et al., 2019; Garcia and Fernandez, 2015; Yu et al., 2019). The approach of (Junges et al., 2016) also requires a finite state space.

The closest study to this paper is the results of (Devonport et al., 2020), where the authors use the results of (Srinivas et al., 2009) to evaluate their confidence in learnt Gaussian processes modelling the system and the environment. However, Devonport et al. do not provide computationally friendly methods for ensuring safety as their framework relies on Lyapunov functions that can be computationally difficult to find (Parrilo, 2000).

2 Safe Learning in the Presence of Gaussian Uncertainty

We consider a discrete time nonlinear control-affine system

$$x_{t+1} = f(x_t) + g(x_t)u_t + w_t, \tag{1}$$

where $x_t \in \mathbb{R}^n$ is the state, $u_t \in \mathbb{R}^m$ is the control input, and $w_t \in \mathbb{R}^n$ is a sequence of independent and identically distributed Gaussian random vectors with zero mean and covariance $\Sigma_w \succeq 0$.[1] See Sections 3 and 4 for extensions of our results to non-zero mean and piecewise constant Gaussian uncertainties. Assume that the model dynamics $f(\cdot)$ and $g(\cdot)$ are known, while the covariance matrix Σ_w is unknown and we are aiming to learn it. The setup, i.e., knowing the system dynamics while not knowing the process noise, can be motivated by that we are using an accurately-modelled agent in an unknown environment (e.g., search and rescue mission by expensive robots). Alternatively, when dealing with a not-fully-known agent, we can lump all model uncertainties (e.g., un-modelled dynamics, model mismatch, and parameter errors) into the process noise and attempt to learn the characteristics of the process noise while maneuvering our agent. The first approach is in spirit close to robotics while the second is similar to robust control methodology.

We have safety constraints of the form

$$H_t x_t \leq h_t, \tag{2}$$

where the inequality is element-wise. This means that, at each iteration, parts of the state space are off limits to us. This could be to avoid collision with other agents, or getting captured by the enemy. Not knowing Σ_w hinders our ability to satisfy this constraint, as we do not know how the process noise influences the state in the next time step after implementing the control action. Therefore, we use the state measurements of the system to learn Σ_w. After $t \geq 1$ time steps, the empirical estimate of the covariance is given by

$$\widehat{\Sigma}_w := \frac{1}{t}\sum_{k=0}^{t-1} w_k w_k^\top = \frac{1}{t}\sum_{k=0}^{t-1}(x_{k+1} - f(x_k) - g(x_k)u_k)(x_{k+1} - f(x_k) - g(x_k)u_k)^\top.$$

Note that $\widehat{\Sigma}_w$ is an unbiased estimate of Σ_w:

$$\mathbb{E}\{\widehat{\Sigma}_w\} = \frac{1}{t}\sum_{k=0}^{t-1}\mathbb{E}\{w_k w_k^\top\} = \Sigma_w.$$

We will make the following assumption:

[1] We say that a matrix $X \succeq 0$ if X is positive semi-definite. For matrices X and Y, we say that $X \succeq Y$ if $X - Y$ is positive semi-definite.

Assumption 1. *There exists some $\sigma > 0$ such that $\Sigma_w \preceq \sigma I$.*

We first give a probabilistic bound on the accuracy of the estimated covariance.

Lemma 2. *For all $t \geq 1$,*

$$\mathbb{P}\left\{\Sigma_w \preceq \widehat{\Sigma}_w + \sigma\sqrt{\frac{n(n+1)}{\delta t}}I\right\} \geq 1 - \delta. \tag{3}$$

Proof. Note that $\widehat{\Sigma}_w \sim \mathcal{W}_n(t^{-1}\Sigma_w, t)$ is Wishart distributed. Thus $\mathbb{E}\{\widehat{\Sigma}_w\} = \Sigma_w$. As a result,

$$\begin{aligned}\mathbb{E}\{(\Sigma_w - \widehat{\Sigma}_w)^2\} &= \mathbb{E}\{\hat{\Sigma}_w^2\} - \Sigma_w^2 \\ &= \frac{\mathrm{trace}(\Sigma_w)}{t}\Sigma_w + \frac{t+1}{t}\Sigma_w^2 - \Sigma_w^2 \\ &= \frac{\mathrm{trace}(\Sigma_w)}{t}\Sigma_w + \frac{1}{t}\Sigma_w^2\end{aligned}$$

where the expression for $\mathbb{E}\{\hat{\Sigma}_w^2\}$ can be found in (Letac and Massam, 2004), see in particular the proof of Theorem 4 there. Now, the matrix version of Chebyshev's inequality given in Lemma 15 in Appendix A results in

$$\begin{aligned}\mathbb{P}\{\Sigma_w - \widehat{\Sigma}_w \preceq \varepsilon I\} &\geq 1 - \mathrm{trace}(\mathbb{E}\{(\Sigma_w - \widehat{\Sigma}_w)^2\varepsilon^{-2}\}) \\ &= 1 - \frac{(\mathrm{trace}(\Sigma_w))^2 + \mathrm{trace}(\Sigma_w^2)}{\varepsilon^2 t},\end{aligned}$$

Hence, by Assumption 1:

$$\mathbb{P}\{\Sigma_w - \widehat{\Sigma}_w \preceq \varepsilon I\} \geq 1 - \frac{\sigma^2 n(n+1)}{\varepsilon^2 t}.$$

Setting $\varepsilon^2 = \sigma^2 n(n+1)/(\delta t)$ gives (3). □

Our approach to controller design in this paper is to take an existing nominal control input $\bar{u}_t$,[2] and modify this control slightly at each iteration to ensure safety. We do this by solving an optimisation problem with safety constraints:

$$u_t = \arg\min_u \quad \|u - \bar{u}_t\|, \tag{4a}$$

$$\text{s.t.} \qquad H_{t+1}(f(x_t) + g(x_t)u + w) \leq h_{t+1}, \forall w : w^\top \left(\widehat{\Sigma}_w + \sigma\sqrt{\frac{2n(n+1)}{\delta t}}I\right)^{-1} w \leq \frac{2n}{\delta}. \tag{4b}$$

We have the following result on high probability safety guarantees of the learned-controlled system.

Theorem 3. *Assume that problem (4) is feasible. Then, if we implement the control action u_t, x_{t+1} is safe with probability of at least $1 - \delta$, i.e., $\mathbb{P}\{H_{t+1}x_{t+1} \leq h_{t+1}\} \geq 1 - \delta$.*

Proof. From Lemma 2, we know that

$$\mathbb{P}\left\{\Sigma_w \preceq \widehat{\Sigma}_w + \sigma\sqrt{\frac{2n(n+1)}{\delta t}}I\right\} \geq 1 - \frac{\delta}{2}.$$

If $\Sigma_w \preceq \widehat{\Sigma}_w + \sigma\sqrt{2n(n+1)/(\delta t)}I$, we have

$$\mathbb{P}\left\{w^\top \left(\widehat{\Sigma}_w + \sigma\sqrt{\frac{2n(n+1)}{\delta t}}I\right)^{-1} w \leq \frac{2n}{\delta}\right\} \geq \mathbb{P}\left\{w^\top \Sigma_w^{-1} w \leq \frac{2n}{\delta}\right\}$$

[2]The nominal control can, for instance, be generated by a reinforcement learning algorithm such as (Lillicrap et al., 2016).

$$\geq 1 - \frac{\delta}{2},$$

where the last inequality follows from an application of Markov's inequality for scalar random variables:

$$\begin{aligned}
\mathbb{P}\left\{w^\top \Sigma_w^{-1} w \leq \frac{2n}{\delta}\right\} &\geq 1 - \frac{\mathbb{E}\{w^\top \Sigma_w^{-1} w\}}{2n/\delta} \\
&= 1 - \frac{\mathrm{trace}(\mathbb{E}\{w w^\top \Sigma_w^{-1}\})}{2n/\delta} \\
&= 1 - \frac{\mathrm{trace}(I)}{2n/\delta} \\
&= 1 - \frac{\delta}{2}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\{H_{t+1} x_{t+1} \leq h_{t+1}\} \\
&\quad \geq \mathbb{P}\left\{\Sigma_w \preceq \widehat{\Sigma}_w + \sigma\sqrt{\frac{2n(n+1)}{\delta t}} I \bigwedge w^\top \left(\widehat{\Sigma}_w + \sigma\sqrt{\frac{2n(n+1)}{\delta t}} I\right)^{-1} w \leq \frac{2n}{\delta}\right\} \\
&\quad = 1 - \mathbb{P}\left\{\Sigma_w \succeq \widehat{\Sigma}_w + \sigma\sqrt{\frac{2n(n+1)}{\delta t}} I \bigvee w^\top \left(\widehat{\Sigma}_w + \sigma\sqrt{\frac{2n(n+1)}{\delta t}} I\right)^{-1} w \geq \frac{2n}{\delta}\right\} \\
&\quad \geq 1 - \mathbb{P}\left\{\Sigma_w \succeq \widehat{\Sigma}_w + \sigma\sqrt{\frac{2n(n+1)}{\delta t}} I\right\} - \mathbb{P}\left\{w^\top \left(\widehat{\Sigma}_w + \sigma\sqrt{\frac{2n(n+1)}{\delta t}} I\right)^{-1} w \geq \frac{2n}{\delta}\right\} \\
&\quad \geq 1 - \frac{\delta}{2} - \frac{\delta}{2} \\
&\quad = 1 - \delta.
\end{aligned}$$

This concludes the proof. □

Note that, as it stands, (4) involves infinitely many constraints and is thus not easy to solve numerically. We next provide a computationally-friendly reformulation of (4). To do so, we first present the following lemma.

Lemma 4. *For $W \succeq 0$ and $d \geq 0$, $\{x \in \mathbb{R}^n \,|\, a^\top x + b^\top w \leq c, \forall w : w^\top W w \leq d\} = \{x \in \mathbb{R}^n \,|\, a^\top x \leq c - \sqrt{d}\|W^{-1/2} b\|_2\}$.*

Proof. With the change of variables $\bar{w} = W^{1/2} w$ and $\bar{b} = W^{-1/2} b$, we have $\{x \in \mathbb{R}^n \,|\, a^\top x + b^\top w \leq c, \forall w : w^\top W w \leq d\} = \{x \in \mathbb{R}^n \,|\, a^\top x + \bar{b}^\top \bar{w} \leq c, \forall \bar{w} : \bar{w}^\top \bar{w} \leq d\}$. Then, following the approach of (Ben-Tal et al., 2009, Example 1.3.3, p. 19), we can obtain $\{x \in \mathbb{R}^n \,|\, a^\top x + \bar{b}^\top \bar{w} \leq c, \forall \bar{w} : \bar{w}^\top \bar{w} \leq d\} = \{x \in \mathbb{R}^n \,|\, \sqrt{d}\|\bar{b}\|_2 \leq c - a^\top x\}$. □

With the aid of Lemma 4, we can rewrite (4) as

$$u_t = \arg\min_u \quad \|u - \bar{u}_t\|, \tag{5a}$$

$$\text{s.t.} \quad H_{t+1}(f(x_t) + g(x_t)u) \leq h_{t+1} - d_{t+1}, \tag{5b}$$

where d_{t+1} is a vector whose i-th entry is equal to

$$d_{t+1}^i = \sqrt{\frac{2n}{\delta}} \left\| \left(\widehat{\Sigma}_w + \sigma\sqrt{\frac{2n(n+1)}{\delta t}} I\right)^{1/2} {H_{t+1}^i}^\top \right\|_2. \tag{6}$$

Here, H_{t+1}^i denotes the i-th row of the matrix H_{t+1}. This reformulation has an interpretation which amounts to tightening of the constraints for a system without uncertainties. The tightening is larger at the beginning and becomes smaller as we gather more information.

Remark 5 (Multi-Agent Safe Learning of the Environment)**.** *The problem formulation of this paper investigates the motion of a single robot in an environment. However, in practice, there are often more than one agent. In that case, we need to decouple the robots, while ensuring that they do not collide and can accomplish tasks collaboratively. For this, we can use the methodology of (Wood et al., 2020) to construct time-varying constraints of the form* (2)*, that can ensure collision avoidance and guarantee collaboration for task assignment. Then, we can use the methodology of this paper for safely maneuvering each robot in the environment.*

3 Non-zero mean Gaussian uncertainty

The assumption of zero mean for the uncertainty w_t may not be appropriate when the uncertainty is not purely random noise but a more systematic uncertainty. In this section, we assume that w_t is Gaussian with mean μ_w and covariance Σ_w, with $\mu_w \neq 0$ in general. We can estimate μ_w and Σ_w using

$$\widehat{\mu}_w := \frac{1}{t}\sum_{k=0}^{t-1} w_k = \frac{1}{t}\sum_{k=0}^{t-1}(x_{k+1} - f(x_k) - g(x_k)u_k)$$

and

$$\begin{aligned}\widehat{\Sigma}_w :=& \frac{1}{t-1}\sum_{k=0}^{t-1}(w_k - \widehat{\mu}_w)(w_k - \widehat{\mu}_w)^\top \\ =& \frac{1}{t-1}\sum_{k=0}^{t-1}(x_{k+1} - f(x_k) - g(x_k)u_k - \widehat{\mu}_w)(x_{k+1} - f(x_k) - g(x_k)u_k - \widehat{\mu}_w)^\top.\end{aligned}$$

In contrast to Section 2, here $\widehat{\Sigma}_w$ is an unbiased estimate of Σ_w when we divide by $t-1$ instead of t. This is because we are estimating both the mean and covariance simultaneously.

Lemma 6. *For all $t \geq 1$,*

$$\mathbb{P}\left\{\|\mu_w - \widehat{\mu}_w\|_2^2 \leq \frac{n\sigma}{t\delta}\right\} \geq 1 - \delta. \tag{7}$$

Proof. Note that $\widehat{\mu}_w - \mu_w$ is zero mean Gaussian with covariance $t^{-1}\Sigma_w$. Then

$$\begin{aligned}\mathbb{E}\{\|\mu_w - \widehat{\mu}_w\|_2^2\} &= \operatorname{trace}(\mathbb{E}\{(\mu_w - \widehat{\mu}_w)(\mu_w - \widehat{\mu}_w)^\top\}) \\ &= \operatorname{trace}(t^{-1}\Sigma_w) \\ &\leq \frac{n\sigma}{t}.\end{aligned}$$

From Markov's inequality, we get

$$\mathbb{P}\left\{\|\mu_w - \widehat{\mu}_w\|_2^2 \leq \varepsilon\right\} \geq 1 - \frac{\mathbb{E}\{\|\mu_w - \widehat{\mu}_w\|_2^2\}}{\varepsilon} \geq 1 - \frac{n\sigma}{t\varepsilon}.$$

Selecting $\delta = n\sigma/(t\varepsilon)$ then gives (7). □

Lemma 7. *For all $t \geq 2$,*

$$\mathbb{P}\left\{\Sigma_w \preceq \widehat{\Sigma}_w + \sigma\sqrt{\frac{n(n+1)}{\delta(t-1)}}I\right\} \geq 1 - \delta. \tag{8}$$

Proof. Note now that $\widehat{\Sigma}_w \sim \mathcal{W}_n((t-1)^{-1}\Sigma, t-1)$ is Wishart distributed (Anderson, 2003, Corollary 7.2.3). The proof then follows by using the same arguments as in the proof of Lemma 2.

□

Letting $\bar{u}_t$ denote a nominal control input, we can then solve

$$u_t = \arg\min_u \quad \|u - \bar{u}_t\|, \tag{9a}$$

$$\text{s.t.} \qquad H_{t+1}(f(x_t) + g(x_t)u + w) \le h_{t+1},\ \forall w \in \Omega, \tag{9b}$$

where

$$\begin{aligned}\Omega :=& \bigcup_{a:\|a\|_2^2 \le 3n\sigma/(t\delta)} \left\{ w \,|\, (w - \widehat{\mu}_w - a)^\top \left(\widehat{\Sigma}_w + \sigma \sqrt{\frac{3n(n+1)}{\delta(t-1)}} I \right)^{-1} (w - \widehat{\mu}_w - a) \le \frac{3n}{\delta} \right\} \\ =& \left\{ w \,|\, (w - \widehat{\mu}_w - a)^\top \left(\widehat{\Sigma}_w + \sigma \sqrt{\frac{3n(n+1)}{\delta(t-1)}} I \right)^{-1} (w - \widehat{\mu}_w - a) \le \frac{3n}{\delta} \text{ for some } a^\top a \le \frac{3n\sigma}{t\delta} \right\}.\end{aligned}$$

Theorem 8. *Assume that problem* (9) *is feasible. Then, if we implement the control action u_t, x_{t+1} is safe with probability of at least $1-\delta$, i.e., $\mathbb{P}\{H_{t+1}x_{t+1} \le h_{t+1}\} \ge 1-\delta$.*

Proof. From Lemma 7, we know that $\mathbb{P}\left\{\Sigma_w \preceq \widehat{\Sigma}_w + \sigma\sqrt{3n(n+1)/(\delta(t-1))}I\right\} \ge 1-\delta/3$, and from Lemma 6, that $\mathbb{P}\left\{\|\mu_w - \widehat{\mu}_w\|_2^2 \le 3n\sigma/(t\delta)\right\} \ge 1-\delta/3$. If $\Sigma_w \preceq \widehat{\Sigma}_w + \sigma\sqrt{3n(n+1)/(\delta(t-1))}I$ and $\|\mu_w - \widehat{\mu}_w\|_2^2 \le 3n\sigma/(t\delta)$, we get

$$\begin{aligned}&\left\{ w \,|\, (w - \mu_w)^\top \Sigma_w^{-1} (w - \mu_w) \le \frac{3n}{\delta} \right\} \\ &\quad\subseteq \left\{ w \,|\, (w - \mu_w)^\top \left(\widehat{\Sigma}_w + \sigma \sqrt{\frac{3n(n+1)}{\delta(t-1)}} I \right)^{-1} (w - \mu_w) \le \frac{3n}{\delta} \right\} \\ &\quad\subseteq \bigcup_{a:\|a\|_2^2 \le 3n\sigma/(t\delta)} \left\{ w \,|\, (w - \widehat{\mu}_w - a)^\top \left(\widehat{\Sigma}_w + \sigma \sqrt{\frac{3n(n+1)}{\delta(t-1)}} I \right)^{-1} (w - \widehat{\mu}_w - a) \le \frac{3n}{\delta} \right\} \\ &\quad= \left\{ w \,|\, \exists a : a^\top a \le \frac{3n\sigma}{t\delta} \bigwedge (w - \widehat{\mu}_w - a)^\top \left(\widehat{\Sigma}_w + \sigma \sqrt{\frac{3n(n+1)}{\delta(t-1)}} I \right)^{-1} (w - \widehat{\mu}_w - a) \le \frac{3n}{\delta} \right\}.\end{aligned}$$

Therefore,

$$\begin{aligned}&\mathbb{P}\left\{ w \,|\, \exists a : a^\top a \le \frac{3n\sigma}{t\delta} \bigwedge (w - \widehat{\mu}_w - a)^\top \left(\widehat{\Sigma}_w + \sigma \sqrt{\frac{3n(n+1)}{\delta(t-1)}} I \right)^{-1} (w - \widehat{\mu}_w - a) \le \frac{3n}{\delta} \right\} \\ &\qquad\qquad \ge \mathbb{P}\left\{ w \,|\, (w - \mu_w)^\top \Sigma_w^{-1} (w - \mu_w) \le \frac{3n}{\delta} \right\} \ge 1 - \frac{\delta}{3}.\end{aligned}$$

Combining these probability bounds results in

$$\begin{aligned}&\mathbb{P}\{H_{t+1}x_{t+1} \le h_{t+1}\} \\ &\ge \mathbb{P}\Bigg\{ \Sigma_w \preceq \widehat{\Sigma}_w + \sigma \sqrt{\frac{3n(n+1)}{\delta(t-1)}} I \bigwedge \|\mu_w - \widehat{\mu}_w\|_2^2 \le \frac{3n\sigma}{t\delta} \bigwedge \Bigg[\exists a : a^\top a \le \frac{3n\sigma}{t\delta} \\ &\qquad\qquad \bigwedge (w - \widehat{\mu}_w - a)^\top \left(\widehat{\Sigma}_w + \sigma \sqrt{\frac{3n(n+1)}{\delta(t-1)}} I \right)^{-1} (w - \widehat{\mu}_w - a) \le \frac{3n}{\delta} \Bigg] \Bigg\} \\ &\ge 1 - \frac{\delta}{3} - \frac{\delta}{3} - \frac{\delta}{3} = 1 - \delta.\end{aligned}$$

This concludes the proof. ∎

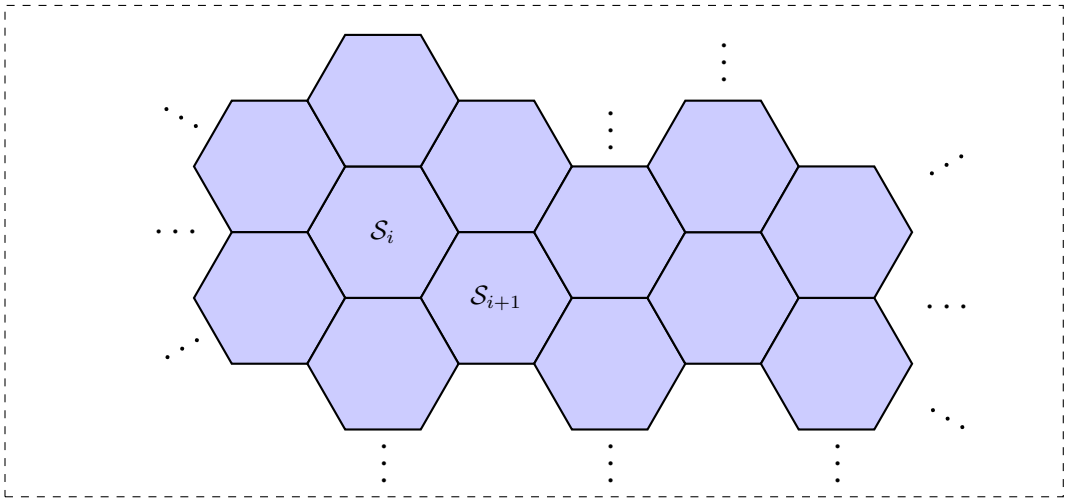


Figure 1: Partitioning the space into sufficiently small non-overlapping cells in which the mean and the covariance of the process noise is approximately constant.

4 Gaussian Uncertainty with Piecewise Constant Mean and Covariance

In practice, the mean and the covariance are rarely fixed across the environment. To overcome this issue, we approximate location-dependent mean $\mu_w(x)$ and covariance $\Sigma_w(x)$ by piecewise constant functions. That is, we partition the space into sufficiently small non-overlapping connected sets $\mathcal{S}_i$ and, in each set, we assume that the mean and the covariance are constant. Therefore, if $x_k \in \mathcal{S}_i$, the process noise is Gaussian with mean $\mu_w^{(i)}$ and covariance $\Sigma_w^{(i)}$. Similar to the earlier sections, we do not know the exact values of the mean $\mu_w^{(i)}$ and covariance $\Sigma_w^{(i)}$ when deploying the agent. Figure 1 illustrates this partitioning.

In each $\mathcal{S}_i$, based on all the measurements in all time instances (up to time t) that we have been inside this region $\forall k : x_k \in \mathcal{S}_i$, we can learn the mean and the covariance of the process noise according to

$$\begin{aligned}
\widehat{\mu}_w^{(i)} :=& \frac{1}{t_i} \sum_{k:x_k \in \mathcal{S}_i} (x_{k+1} - f(x_k) - g(x_k)u_k), \\
\widehat{\Sigma}_w^{(i)} :=& \frac{1}{t_i - 1} \sum_{k:x_k \in \mathcal{S}_i} (x_{k+1} - f(x_k) - g(x_k)u_k - \widehat{\mu}_w^{(i)})(x_{k+1} - f(x_k) - g(x_k)u_k - \widehat{\mu}_w^{(i)})^\top,
\end{aligned}$$

where t_i is the number of measurements in region $\mathcal{S}_i$:

$$t_i := \sum_{k=0}^{t-1} \mathbb{1}_{x_k \in \mathcal{S}_i}.$$

Note that we do not need to keep a record of all visited locations to implement these estimators. We can use an online estimator of the form:

$$\begin{aligned}
t_i &\leftarrow t_i + 1, \\
\widehat{\mu}_w^{(i)} &\leftarrow \frac{t_i - 1}{t_i} \widehat{\mu}_w^{(i)} + \frac{1}{t_i}(x_{k+1} - f(x_k) - g(x_k)u_k) \\
\widehat{\Sigma}_w^{(i)} &\leftarrow \frac{t_i - 2}{t_i - 1} \widehat{\Sigma}_w^{(i)} + \frac{1}{t_i - 1}(x_{k+1} - f(x_k) - g(x_k)u_k - \widehat{\mu}_w^{(i)})(x_{k+1} - f(x_k) - g(x_k)u_k - \widehat{\mu}_w^{(i)})^\top
\end{aligned}$$

if $x_k \in \mathcal{S}_i$. Therefore, we only need to store $n(n+1)/2$ elements of $\widehat{\Sigma}_w^{(i)}$, n elements of $\widehat{\mu}_w^{(i)}$, and t_i, totalling $n^2/2 + 3n/2 + 1$ scalars for each region.

Similar to Lemmas 6 and 7, we have the following:

Corollary 9. *The following hold:*

$$\mathbb{P}\left\{\|\mu_w^{(i)} - \widehat{\mu}_w^{(i)}\|_2^2 \leq \frac{n\sigma}{t_i\delta}\right\} \geq 1-\delta, \quad \forall t_i \geq 1, \tag{10}$$

$$\mathbb{P}\left\{\Sigma_w^{(i)} \preceq \widehat{\Sigma}_w^{(i)} + \sigma\sqrt{\frac{n(n+1)}{\delta(t_i-1)}}I\right\} \geq 1-\delta, \quad \forall t_i \geq 2. \tag{11}$$

Therefore, if $x_k \in S_i$, we can solve the robust optimization (9) to manoeuvre safely. The bounds in Corollary 9 may sometimes be conservative, since they do not use the fact that we have gathered measurements in a surrounding region $\mathcal{S}_j$, and that the changes in the mean $\mu_w^{(i)}$ and covariance $\Sigma_w^{(i)}$ cannot be arbitrarily abrupt. We say that $\mathcal{S}_j$ is a *neighbour* of $\mathcal{S}_i$ if there exists a control action that can move an agent from some point in $\mathcal{S}_i$ to another point in $\mathcal{S}_{i+1}$ in one time step. For instance, in Figure 1, regions $\mathcal{S}_i$ and $\mathcal{S}_{i+1}$ are neighbours.

Assumption 10. *For any two neighbouring regions $\mathcal{S}_i$ and $\mathcal{S}_j$, $\|\mu_w^{(i)} - \mu_w^{(j)}\| \leq \rho_1$ and $\|\Sigma_w^{(i)} - \Sigma_w^{(j)}\| \leq \rho_2$ for some constants $\rho_1, \rho_2 > 0$.*

Remark 11. *Instead of the norm bounds in Assumption 10, we can use any metric on probability distributions, such as the Wasserstein distance (Givens and Shortt, 1984).*

Lemma 12. *Let $\mathcal{N}_i$ denote any set consisting of the indices of neighbours of $\mathcal{S}_i$ for which $t_j \geq 1$ for all $j \in \mathcal{N}_i$. Denote $m_i = |\mathcal{N}_i|$. Then,*

$$\mathbb{P}\left\{\|\mu_w^{(i)} - \widehat{\mu}_w^{(i)}\|_2^2 \leq \frac{n\sigma(m_i+1)}{t_i\delta} \bigwedge \|\mu_w^{(i)} - \widehat{\mu}_w^{(j)}\|_2^2 \leq \left(\rho_1 + \sqrt{\frac{n\sigma(m_i+1)}{t_j\delta}}\right)^2, \forall j \in \mathcal{N}_i\right\} \geq 1-\delta.$$

Proof. Note that, for all $j \in \mathcal{N}_i$, $\mathbb{P}\left\{\|\mu_w^{(j)} - \widehat{\mu}_w^{(j)}\|_2^2 \leq n\sigma(m_i+1)/(t_j\delta)\right\} \geq 1 - \delta/(m_i+1)$ and $\|\mu_w^{(i)} - \mu_w^{(j)}\| \leq \rho_1$. Thus,

$$\begin{aligned}
&\mathbb{P}\left\{\|\mu_w^{(i)} - \widehat{\mu}_w^{(j)}\|_2^2 \leq \left(\rho_1 + \sqrt{\frac{n\sigma(m_i+1)}{t_j\delta}}\right)^2\right\} \\
&\qquad \geq \mathbb{P}\left\{\|\mu_w^{(i)} - \mu_w^{(j)}\|_2 + \|\mu_w^{(j)} - \widehat{\mu}_w^{(j)}\|_2 \leq \rho_1 + \sqrt{\frac{n\sigma(m_i+1)}{t_j\delta}}\right\} \\
&\qquad \geq \mathbb{P}\left\{\|\mu_w^{(j)} - \widehat{\mu}_w^{(j)}\|_2 \leq \sqrt{\frac{n\sigma(m_i+1)}{t_j\delta}}\right\} \geq 1 - \frac{\delta}{m_i+1}.
\end{aligned}$$

Combining these probability inequalities with

$$\mathbb{P}\left\{\|\mu_w^{(i)} - \widehat{\mu}_w^{(i)}\|_2^2 \leq \frac{n\sigma(m_i+1)}{t_i\delta}\right\} \geq 1 - \frac{\delta}{m_i+1}$$

concludes the proof. □

The bound in Lemma 12 can be tighter than the generic bound (10) in Corollary 9, depending on the values of m_i and t_j for $j \in \mathcal{N}_i$.

Lemma 13. *Let $\mathcal{N}_i$ denote any set consisting of the indices of neighbours of $\mathcal{S}_i$ for which $t_j \geq 2$ for all $j \in \mathcal{N}_i$. Denote $m_i = |\mathcal{N}_i|$. Then,*

$$\begin{aligned}
\mathbb{P}\Bigg\{&\Sigma_w^{(i)} \preceq \widehat{\Sigma}_w^{(i)} + \sigma\sqrt{\frac{n(n+1)(m_i+1)}{\delta(t_i-1)}}I \\
&\bigwedge\left[\Sigma_w^{(j)} \preceq \widehat{\Sigma}_w^{(j)} + \sigma\sqrt{\frac{n(n+1)(m_i+1)}{\delta(t_j-1)}}I \bigwedge \|\Sigma_w^{(i)} - \Sigma_w^{(j)}\| \leq \rho_2\right], \forall j \in \mathcal{N}_i\Bigg\} \geq 1-\delta.
\end{aligned}$$

Proof. The proof follows from a similar reasoning as in the proof of Lemma 12. □

5 Conclusions

We focused on guaranteeing safety while learning and control proceed simultaneously. We considered the problem of safely navigating a nonlinear control-affine system subject to unknown additive uncertainties. We modelled uncertainty as an additive Gaussian noise with unknown mean and covariance. Subsequently, we used state measurements to learn the mean and covariance of the process noise. We provided rigorous time-varying confidence intervals on the empirical mean and covariance of the uncertainty. This allowed us to employ the learned moments of the Gaussian noise along with the mentioned confidence bounds to modify the control input via a robust optimisation program with safety constraints encoded in it. We proved that we can guarantee that the state will remain in the safe set with an arbitrarily large probability while learning and control are carried out simultaneously. We provided a secondary formulation based on tightening the safety constraints to counter the uncertainty about the learned mean and variance. The magnitude of the tightening can be decreased as our confidence in the learned mean and variance increases. Finally, noting that in most realistic cases the environment and its effect on our agent changes within the space, we extend our framework to admit Gaussian uncertainties with piecewise constant mean and covariance.

Future work can focus on a number of important avenues. First, the optimization problem in (9) currently has infinitely many constraints. A future step is to rewrite this problem in terms of constraint tightening using robust optimization. This makes the problem computationally tractable. Second, Section 4 considers a piecewise constant mean and covariance noise model. Another future direction is to learn more general nonlinear noise models, for instance a polynomial noise model. Finally, the current results are theoretical and will be validated in the future on a real robot in an unknown environment.

References

A. D. Ames, J. W. Grizzle, and P. Tabuada. Control barrier function based quadratic programs with application to adaptive cruise control. In *Proc. IEEE Conf. Decision and Control*, pages 6271–6278, Los Angeles, USA, December 2014.

A. D. Ames, X. Xu, J. W. Grizzle, and P. Tabuada. Control barrier function based quadratic programs for safety critical systems. *IEEE Trans. Autom. Control*, 62(8):3861–3876, August 2017.

T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics. Wiley, 3rd edition, 2003.

A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton Series in Applied Mathematics. Princeton University Press, 2009.

R. Bhatia. *Positive Definite Matrices*. Princeton Series in Applied Mathematics. Princeton University Press, 2015.

R. Cheng, G. Orosz, R. M. Murray, and J. W. Burdick. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In *Proc. AAAI-19*, Honolulu, USA, January 2019.

R. Cheng, M. J. Khojasteh, A. D. Ames, and J. W. Burdick. Safe multi-agent interaction through robust control barrier functions with learned uncertainties. arXiv preprint arXiv:2004.05273, 2020.

J. Choi, F. Castaneda, C. J. Tomlin, and K. Sreenath. Reinforcement learning for safety-critical control under model uncertainty, using control Lyapunov functions and control barrier functions. arXiv preprint arXiv:2004.07584, 2020.

A. Devonport, H. Yin, and M. Arcak. Bayesian safe learning and control with sum-of-squares analysis and polynomial kernels. In *Proc. IEEE Conf. Decision and Control*, pages 3159–3165, Jeju Island, South Korea, December 2020.

J. Garcia and F. Fernandez. A comprehensive survey on safe reinforcement learning. *J. Machine Learning Research*, 16:1437–1480, August 2015.

C. R. Givens and R. M. Shortt. A class of Wasserstein metrics for probability distributions. *The Michigan Mathematical Journal*, 31(2):231–240, 1984.

P. Jagtap, G. J. Pappas, and M. Zamani. Control barrier functions for unknown nonlinear systems using Gaussian processes. arXiv preprint arXiv:2010.05818, 2020.

S. Junges, N. Jansen, C. Dehnert, U. Topcu, and J.-P. Katoen. Safety-constrained reinforcement learning for MDPs. In *Proc. TACAS*, pages 130–146, Eindhoven, The Netherlands, April 2016.

S. Kolathaya and A. D. Ames. Input-to-state safety with control barrier functions. *IEEE Control Systems Letters*, 3(1):108–113, January 2019.

G. Letac and H. Massam. All invariant moments of the Wishart distribution. *Scand. J. Statistics*, 31(2): 295–318, June 2004.

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. In *Proc. ICLR*, San Juan, Puerto Rico, May 2016.

H. Lütkepohl. *Handbook of Matrices*. Wiley, 1996.

P. A. Parrilo. *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*. PhD thesis, California Institute of Technology, 2000.

N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

A. J. Taylor, A. Singletary, Y. Yue, and A. D. Ames. Learning for safety-critical control with control barrier functions. In *Proc. Conf. Learning for Dynamics and Control*, 2020.

T. A. Wood, M. Khoo, E. Michael, C. Manzie, and I. Shames. Collision avoidance based on robust lexicographic task assignment. *IEEE Robotics and Automation Letters*, 5(4):5693–5700, 2020.

M. Yu, Z. Yang, M. Kolar, and Z. Wang. Convergent policy optimization for safe reinforcement learning. In *Proc. NeurIPS*, Vancouver, Canada, December 2019.

W. Zhang, O. Bastani, and V. Kumar. MAMPS: Safe multi-agent reinforcement learning via model predictive shielding. arXiv preprint arXiv:1910.12639, 2019.

A Matrix concentration inequality

Lemma 14 (Matrix Markov's inequality)**.** *Let $A \succ 0$, and let X be a random matrix such that $X \succeq 0$ almost surely. Then,* $\mathbb{P}\{X \preceq A\} \geq 1 - \text{trace}(\mathbb{E}\{X\}A^{-1})$.

Proof. We prove the reverse $\mathbb{P}\{X \not\preceq A\} \leq \text{trace}(\mathbb{E}\{X\}A^{-1})$. For this, we need to show that $\mathbb{1}_{X \not\preceq A} \leq \text{trace}(A^{-1/2}XA^{1/2})$, where $\mathbb{1}$ is an indicator function, i.e., $\mathbb{1}_p$ is equal to one if p holds and zero otherwise. Note that $\text{trace}(A^{-1/2}XA^{1/2}) \geq 0$ because $X \succeq 0$. It remains to show that $\text{trace}(A^{-1/2}XA^{1/2}) \geq 1$ if $X \not\preceq A$. Assume that $X \not\preceq A$. There must exist $v \neq 0$ such that $v^\top(A - X)v < 0$. Define $w = A^{1/2}v$. We get $w^\top w - w^\top A^{-1/2}XA^{-1/2}w < 0$. Therefore, $w^\top A^{-1/2}XA^{-1/2}w/w^\top w > 1$. Note that $\lambda_{\max}(A^{-1/2}XA^{-1/2}) = \max_{y \neq 0} y^\top A^{-1/2}XA^{-1/2}y/y^\top y \geq w^\top A^{-1/2}XA^{-1/2}w/w^\top w > 1$, where the equality follows from Rayleigh–Ritz Theorem (Lütkepohl, 1996, Item (1),§5.2.2). Furthermore, since $A^{-1/2}XA^{-1/2} \succeq 0$ and thus all its eigenvalues are positive, we get $\text{trace}(A^{-1/2}XA^{-1/2}) \geq \lambda_{\max}(A^{-1/2}XA^{-1/2}) > 1$. Finally, note that $\mathbb{P}\{X \not\preceq A\} = \mathbb{E}\{\mathbb{1}_{X \not\preceq A}\} \leq \mathbb{E}\{\text{trace}(A^{-1/2}XA^{1/2})\} = \text{trace}(\mathbb{E}\{X\}A^{-1})$. □

Lemma 15 (Matrix Chebyshev's inequality)**.** *Let $A \succ 0$, and let X be a random matrix such that $X \succeq 0$ almost surely. Then,* $\mathbb{P}\{X \preceq A\} \geq 1 - \text{trace}(\mathbb{E}\{X^2\}A^{-2})$.

Proof. First, note that $\{X \succeq 0 | X^2 \preceq A^2\} \subseteq \{X \succeq 0 | X \preceq A\}$ because $X^2 \preceq A^2$ implies $X \preceq A$ (Bhatia, 2015, Proposition 1.2.9). Therefore, $\mathbb{P}\{X \preceq A\} \geq \mathbb{P}\{X^2 \preceq A^2\}$. The rest follows from an application of Lemma 14 on $\mathbb{P}\{X^2 \preceq A^2\}$. □